Protein inference based on peptides identified from tandem mass spectra

A Thesis Submitted to the

College of Graduate Studies and Research

in Partial Fulfillment of the Requirements

for the degree of Doctor of Philosophy

in the Division of Biomedical Engineering

University of Saskatchewan

Saskatoon

By

Jinhong Shi

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Division of Biomedical Engineering

57 Campus Dr.

University of Saskatchewan

Saskatoon, Saskatchewan

Canada

S7N 5A9

Abstract

Protein inference is a critical computational step in the study of proteomics. It lays the foundation for further structural and functional analysis of proteins, based on which new medicine or technology can be developed. Today, mass spectrometry (MS) is the technique of choice for large-scale inference of proteins in proteomics. In MS-based protein inference, three levels of data are generated: (1) tandem mass spectra (MS/MS); (2) peptide sequences and their scores or probabilities; and (3) protein sequences and their scores or probabilities. Accordingly, the protein inference problem can be divided into three computational phases: (1) process MS/MS to improve the quality of the data and facilitate subsequent peptide identification; (2) postprocess peptide identification results from existing algorithms which match MS/MS to peptides; and (3) infer proteins by assembling identified peptides. The addressing of these computational problems consists of the main content of this thesis.

The processing of MS/MS data mainly includes denoising, quality assessment, and charge state determination. Here, we discuss the determination of charge states from MS/MS data using low-resolution collision induced dissociation. Such spectra with multiple charges are usually searched multiple times by assuming each possible charge state. Not only does this strategy increase the overall database search time, but also yields more false positives. Hence, it is advantageous to determine the charge states of such spectra before the database search. A new approach is proposed to determine the charge states of low-resolution MS/MS. Four novel and discriminant features are adopted to describe each MS/MS and are used in Gaussian mixture model to distinguish doubly and triply charged peptides. The results have shown that this method can assign charge states to low-resolution MS/MS more accurately than existing methods.

Many search engines are available for peptide identification. However, there is usually a high false positive rate (FPR) in the results. This can bring many false identifications to protein inference. As a result, it is necessary to postprocess peptide identification results. The most commonly used method is performing

statistical analysis, which does not only make it possible to compare and combine the results from different search engines, but also facilitates subsequent protein inference. We proposed a new method to estimate the accuracy of peptide identification with logistic regression (LR) and exemplify it based on Sequest scores. Each peptide is characterized with the regularized Sequest scores ΔCn^* and $Xcorr^*$. The score regularization is formulated as an optimization problem by applying two assumptions: the smoothing consistency between sibling peptides and the fitting consistency between original scores and new scores. The results have shown that the proposed method can robustly assign accurate probabilities to peptides and has a very high discrimination power, higher than that of PeptideProphet, to distinguish correctly and incorrectly identified peptides.

Given identified peptides and their probabilities, protein inference is conducted by assembling these peptides. Existing methods to address this MS-based protein inference problem can be classified into two groups: two-stage and one unified framework to identify peptides and infer proteins. In two-stage methods, protein inference is based on, but also separated from, peptide identification. Whereas in one unified framework, protein inference and peptide identification are integrated together. In this study, we proposed a unified framework for protein inference, and developed an iterative method accordingly to infer proteins based on Sequest peptide identification. The statistical analysis of peptide identification is performed with the LR previously introduced. Protein inference and peptide identification are iterated in one framework by adding a feedback from protein inference to peptide identification. The feedback information is a list of high-confidence proteins, which is used to update the adjacency matrix between peptides. The adjacency matrix is used in the regularization of peptide scores. The results have shown that the proposed method can infer more true positive proteins, while outputting less false positive proteins than ProteinProphet at the same FPR. The coverage of inferred proteins is also significantly increased due to the selection of multiple peptides for each MS/MS spectrum and the improvement of their scores by the feedback from the inferred proteins.

Acknowledgements

I would like to thank my supervisor Prof. Fang-Xiang Wu. His expertise, understanding and patience have always been my great support during the course of my study and research. Throughout my entire research, he provided me with constant encouragement, sound advice, and many research ideas and directions. I appreciate his knowledge, skill, and his passion for research. It has been a great honor to finish my graduate study under his supervision.

My sincere thanks also go to other members of my advisory committee, Prof. Chris Zhang, Prof. Randy Purves and Prof. Ian McQuillan, for their assistance and great advice they provided during my PhD program.

I also want to thank my group members, Jiarui Ding, Lei Mu, Zheng Yuan, Wenjun Lin, Bolin Chen, Yan Yan, Lizhi Liu, and Vivian Fan, for their help in my life and their advice in my research work. They all make good friends.

I would also like to thank my family for their love and support through my entire life. Also, my thanks go to all my friends in Saskatoon. Without them, my stay in Saskatoon would not be having so much fun.

Finally, I gratefully acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) and College of Graduate Studies and Research at the University of Saskatchewan for their financial supports of my studies.

This thesis is dedicated to my family.

Contents

List of Tables

List of Figures

List of Abbreviations

AUC	Area Under the Curve
CID	Collision Induced Dissociation
ESI	ElectroSpray Ionization
EM	Expectation-Maximization
FPR	False Positive Rate
FDR	False Discovery Rate
GMM	Gaussian Mixture Model
HPLC	High Pressure Liquid Chromatography
IEF	IsoElectric Focusing
LC	Liquid Chromatography
LIT	Linear Ion Trap
LR	Logistic Regression
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
MALDI	Matrix Assisted Laser Desorption/Ionization
NSP	Number of Sibling Peptides
PPI	Protein-Protein Interaction
PSM	Peptide-Spectrum-Match
PMF	Peptide Mass Fingerprinting
pI	Isoelectric Point
PTM	Post Translational Modification
ROC	Receiver Operating Characteristic
TOF	Time of Flight

Chapter 1

Introduction

1.1 Overview of the protein inference problem

Proteomics is the large-scale study of proteins expressed in a sample that is extracted from a tissue or an organism. This study covers much of the functional analysis of gene products, including the identification and characterization of proteins, and protein-protein interactions (PPI) [1]. It can provide complementary information which cannot be provided by the study of genomics or transcriptomics. One of the explicit aims of proteomics is to infer proteins in a cell or tissue, or eventually in a whole organism. Therefore, protein inference is an important step in proteomics, which is referred to as assembling identified peptides to infer the protein content in a biological sample [2]. Currently, mass spectrometry (MS) is the technique of choice to accomplish this goal [3–5]. The general steps of this MS-based shotgun proteomics for protein inference are shown in Figure 1.1. First, proteins are digested into smaller peptides with enzymes. Then, tandem mass spectrometry (MS/MS) spectra are obtained from a combination of liquid chromatography (LC) and mass spectrometry. Next, peptide identification is performed by database searching or de novo sequencing. Since peptide identification is usually a large-scale analysis, and there is a high rate of false positive identifications, postprocessing peptide identification results is necessary to ensure the quality of peptide identification. The most commonly used method is to perform a statistical analysis of peptide identification results. Finally, protein inference is conducted by assembling the identified peptides with the assistance of available protein

databases.

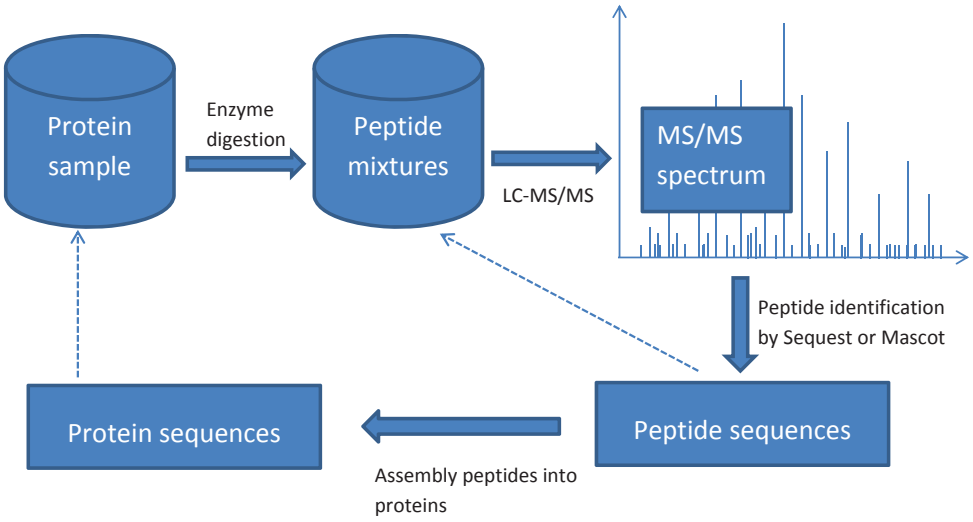


Figure 1.1: The general experimental steps in the shotgun proteomics for protein inference.

As mentioned above, in shotgun proteomics, proteins are first digested into peptides, and protein inference is conducted by identifying peptides first and then assembling these peptides to obtain final proteins. We are more interested in which proteins are contained in a sample, and peptide identification is a necessary intermediate step in protein inference based on shotgun proteomics. After gathering all identified peptides, we need to infer the existence of proteins in the sample. The natural nested relationship between identified peptides and proteins in a database can be represented with a bipartite graph, in which there are only edges between nodes in the upper level and nodes in the lower level, whereas there are no edges between nodes in the same level. Figure 1.2 gives a typical example of the relationship between peptides and proteins. In most cases, this is the standard input for the protein inference model. One common problem of inferring proteins from such bipartite graphs is the existence of degenerate peptides that are shared by multiple proteins in a database. More details about the degenerate peptides will be discussed later. Here, we first see a simple example in Figure 1.2, where peptide P_4 is shared by protein Q_A, protein Q_B and protein Q_C. If there is no other supporting information, it is hard to decide to which protein should peptide P_4 be assigned. In addition, 'one-hit wonders' also commonly happen in protein inference. As shown in Figure 1.2, protein

Q_D is a 'one-hit wonder'. Even if peptide P_5 is unique to protein Q_D, it is still unreliable to determine the presence of protein Q_D in the sample, because there is a chance that peptide P_5 itself is a false positive.

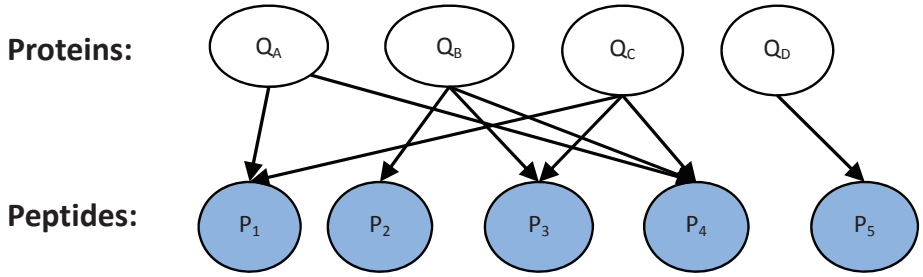


Figure 1.2: A typical bipartite graph which shows the relationship between identified peptides and database proteins.

Aside from the two relatively special occasions, degenerate peptides and 'one-hit wonder', discussed above, the accuracy of peptide identification can affect protein inference in a more general and broad sense. Usually, protein inference models are built by making some necessary assumptions. First, it is assumed that all identified peptides that are used for protein inference are true positives. From this, we can derive the upper and lower bound of the number of possible proteins in a sample. Obviously, the upper bound is the number of all proteins which are associated with the identified peptides. This upper bound sets the limit, and we cannot infer more proteins based on the current input. Some but few existing protein inference methods return all possible proteins without filtering [6], and the underlying assumption is that the sample of interest contains a large portion of homologous proteins. It is not as simple as it seems to find the lower bound of the number of possible proteins. This question can be formulated as a set covering problem [7], of which the goal is to find a minimum subset of proteins that cover all the identified peptides. Then, the lower bound is the number of proteins in the optimal solution of the set covering problem. It is well known that the set covering problem is NP-complete, which makes it difficult to obtain the optimal solution in practice [8]. Therefore, parsimony principle is usually used, which applies Occam's razor [9] to deal with degenerate peptides. According to this principle, only the simplest group of proteins which are sufficient to explain all the observed peptides are reported to be identified [10, 11]. In summary, under this problem setting, any

protein inference algorithm can only produce a result in between the upper and lower bound, trying to reach a trade-off between not including too many false positive proteins and not excluding too many true positive proteins. On the one hand, reporting the upper bound number of proteins may include too many false positive proteins in the final result. Noted that the upper bound is deducted by searching the database with the identified peptides, and thus it is only a theoretical upper bound based on the given identified peptides and the protein database. On the other hand, reporting the lower bound number of proteins may exclude some true positive proteins, especially in the case of homologous proteins existing in the sample.

The above derivation of the upper and lower bound of the number of proteins are based on the given assumption as well as that no other information is adopted in protein inference. If the given identified peptides are not considered all true, then the lower bound may be even lower. Similarly, if we take advantage of some supplementary information such as, raw MS/MS data, single-stage MS data, peptide expression profiles, mRNA expression data, PPI networks or gene models, to assist the inference of proteins, then the upper bound can be raised. Nowadays, there is a trend that protein inference is performed by combining MS/MS data with other available information in order to increase the number of inferred proteins [7, 12–18].

In addition, it is also usually presumed that all peptides have the same chance to be detected in an MS experiment. As is known, some peptides have a higher chance to be detected than others in the same experimental conditions. The reasons behind this include the difference between the ability of peptides to be ionized and fragmented in mass spectrometers, and the difference between the physio-chemical properties of their parent proteins in the enzymatic digestion. This is why the concept of proteotypic peptides, which is referred to as the peptides in a protein that are most likely to be observed by current MS-based proteomics method [19, 20], was proposed and has been widely used in quantitative proteomics. Besides, peptide detectability, which is defined as the probability of observing a peptide in a standard sample by a standard proteomics routine [21], was also proposed to address the problem of the assignment of degenerate peptides.

Over the past decade, much attention was given to the peptide identification from MS/MS data, which

includes developing algorithms to match MS/MS data to peptides [22–27], and postprocessing the peptide identification results from those algorithms [28–35]. However, it is not straightforward to generate a list of confidently inferred proteins from those identified peptides. The reasons are given as follows. First, as discussed previously, the assembly of peptides to corresponding proteins is complicated by the existence of degenerate peptides. The uncertainty of assigning degenerate peptides to truly present proteins brings much ambiguity to protein inference, not only which confounds the identification of truly present protein(s) when it/they indeed exist in the sample, but also increases the number of false positive inferred proteins when the degenerate peptides themselves are false positives. This is a big challenge in protein inference, and some existing attempts to address this problem will be introduced in the next chapter. Secondly, 'one-hit wonders' are often seen in the protein inference process [36], which puts us in a dilemma of keeping or discarding these proteins. Literally, 'one-hit wonders' are proteins that only have one single peptide identified in an MS experiment. On the one hand, we want to increase the number of proteins inferred from MS data in order to improve the coverage of the sample. On the other hand, we want to keep the number of false positive proteins as low as possible. It is difficult to determine the presence of a protein confidently only based on one peptide from it, even if this peptide is unique to this protein in the database since it is possible that this peptide itself is a false positive. Thirdly, it is also a big challenge to validate the results of a protein inference model in proteomics. The most direct method is to use some datasets in which the validity of proteins is known to us in advance. Some standard datasets are already collected for the purpose of verifying algorithms for peptide identification and protein inference [37, 38]. However, such benchmark datasets usually contain very few proteins, which are far from comparable to the complexity of datasets in real proteomics projects. Thus, these standard datasets can only provide a very limited performance assessment and comparison of different models. An alternative way is to generate simulation data that are reasonably close to reality and provide a fair testing ground for different models [39]. The drawback of the simulation datasets is that they completely depend on the underlying assumptions of the generating model, and thus they have inevitable biases which are not expected in the assessment and comparison of performance of different models. Although real and representative reference datasets with ground-truth are desirable in the evaluation of protein inference

results, they are too expensive to generate and collect, especially when other supplementary information besides MS data, such as gene models or PPI networks, is required to assist the inference of proteins. As a result, in most cases, we can only estimate the reliability of protein inference. Overall, protein inference is not as simple as it seems. There are many challenges associating with this problem, and a lot of effort is still needed to address those challenges in order to produce a reliable and close to complete list of proteins.

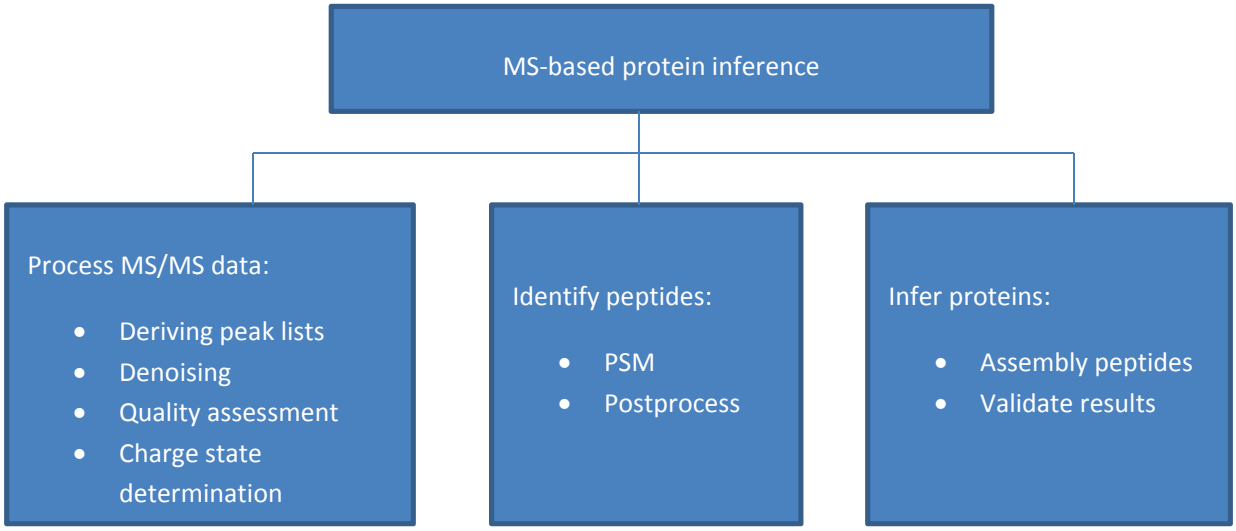


Figure 1.3: Three general computational phases in the MS-based protein inference, which are grouped according to the data subject to be processed. PSM is short for peptide-spectrum-match.

As shown in Figure 1.1, there are three levels of data in MS-based protein inference: MS/MS spectra, identified peptides and inferred proteins. These data correspond to three general computational phases in protein inference: (1) processing MS/MS data; (2) identifying peptides from MS/MS data; and (3) protein inference by assembling identified peptides. These computational steps are grouped and shown in Figure 1.3. In the first phase, MS/MS data are processed to facilitate and improve the analysis in the second phase of peptide identification. Raw MS/MS data are continuous and usually heavily contaminated by noise. Thus, the first step is to transform these continuous data into peak lists with discrete data points consisting of horizontal mass-to-charge (m/z) and vertical intensity. This step is introduced in detail in Chapter 2. After this step, the peak lists are subject to all follow-up processing, which includes denoising, quality

assessment and charge state determination of MS/MS data. Machine learning methods such as support vector machine (SVM) and Gaussian mixture model have been applied to determine the charge states of MS/MS spectra [40, 41], and the k-means clustering and SVM have been used in the quality assessment of MS/MS spectra [42, 43]. Novel features are constructed to describe each tandem mass spectrum, and they are used in the machine learning methods to discriminate, for example, high or low quality MS/MS spectra. After being processed, MS/MS data are more ready and convenient for peptide identification.

In the second phase, peptides are identified by matching MS/MS data to peptide sequences in the database. Since there are already many well-developed search engines performing the work of peptide-spectrum-match (PSM), it is supposed here that PSM has been conducted and the peptide identification reports from search engines are ready for our use. As is known, most search engines provide a group of scores to measure the degree of match between MS/MS and peptide sequences from different angles. Although a group of scores demonstrates a more comprehensive view on one peptide identification, and it helps 'people' to better understand the match between spectrum and peptide, it is not helpful for 'computers' in the same sense. Due to the large scale of modern proteomics analysis, it is almost impractical for people to verify each peptide identification. Under this situation, the statistical analysis which can transform a group of scores into one probability becomes a necessary step in postprocessing peptide identification results from search engines. Furthermore, different engines use different scoring functions, which leads one PSM to have multiple groups of scores. In this case, it is hard to compare or combine the results from different search engines. However, by transforming all scores with statistical analysis into the same scale of probability, the comparison and combination of multiple PSM results can be realized. Besides, statistical analysis can estimate the accuracy of peptide identification and facilitate the subsequent protein inference [2, 10, 28, 33–35, 44]. Generally, novel predictors are proposed based on scores output from search engines, and probabilistic machine learning methods can be used to statistically analyze peptide identification results.

In the third phase, protein inference can be fulfilled by assembling peptides identified in the last step. Based on different criteria, the protein inference model can be categorized into different groups. According to the

data and information used in the model, the protein inference model may be classified into one using only MS data, and the other one using MS data and extra information such as gene models or PPI networks. Although additional information other than MS data is helpful in protein inference, the availability of this kind of information is limited to very few organism models. For MS-based protein inference, existing methods can be further split into two groups. The first group performs protein inference and peptide identification separately [10, 12, 45, 46]. First, peptides are identified from MS/MS data by de novo sequencing [22–24] or database search [25–27]. Then, proteins are inferred by assembling these identified peptides. The other group combines protein inference with peptide identification, identifying peptides and proteins simultaneously. It has been shown that the trend of MS-based protein inference is to unify protein inference and peptide identification in one framework, because this way can make better use of the available information from MS/MS data to inferred proteins [18, 44, 47–49]. After the protein inference model has been developed, it is also necessary to figure out how to evaluate the performance of the model. It is very important to validate the protein inference results in real proteomics projects, the aim of which may be to find effective biomarkers for medicine development.

1.2 Motivation, goals and organization of this thesis

1.2.1 Motivation and goals

As previously mentioned, much attention was given to peptide identification based on MS/MS data in the past decade. Relatively, protein inference is less sufficiently studied compared to the extensive study of peptide identification. In the early research work on protein inference, which includes the very popular program ProteinProphet [10], often ignored is the natural nested relationship between identified peptides and database proteins. In such cases, protein inference and peptide identification are two separate computational steps with the results of peptide identification as the input into protein inference. It has been shown that only

a small portion of MS/MS data can be interpreted by available search engines [38], and furthermore, there is usually a high false positive rate (FPR) in the resultant peptides. Consequently, two obvious problems will happen when protein inference is performed with the input of 'one-step' peptide identification results only based on MS/MS data. One problem is that, for most inferred proteins, the coverage is usually very low due to the small number of peptides identified from MS/MS data. Also, this problem compromises the accuracy of inferred proteins, since the more peptides identified for a protein, the more reliable that this protein is inferred to exist in the sample. The other problem is that there will also be a high FPR of inferred proteins which is attributed by the high FPR of identified peptides, and the FPR of inferred proteins can even be magnified due to the 'one-to-many' mapping relationship between degenerate peptides and their parent proteins.

With the observation of this ignorance of the nested relationship between identified peptides and database proteins in traditional protein inference methods, we are motivated to design a unified framework that can infer proteins and output identified peptides simultaneously based on peptide identification reports from search engines. This framework will integrate protein inference and postprocessing of peptide identification together by allowing a feedback from protein inference to the postprocessing of peptide identification. Inspired by reference [50], which uses the sibling relationship between peptides to regularize the scores of peptides from search engines, we formulate the feedback information, a list of putative inferred proteins, as the construction of an adjacency matrix between peptides. Each element in the matrix takes the values of 1 if two peptides are siblings, or 0 otherwise, if two peptides are not siblings. Two peptides are siblings if they can be generated by a common parent protein. Furthermore, an iterative method is developed accordingly based on the proposed unified framework to infer proteins and identify peptides simultaneously.

As shown in Figure 1.1, there are three levels of data involved in protein inference, and correspondingly, there are three computational phases. In addition to the final goal of inferring proteins by assembling peptides identified from MS/MS data, the processing of MS/MS data and the postprocessing of peptide identification results are also studied as part of the preparation work for protein inference.

1.2.2 Organization of the thesis

This thesis is organized in a manuscript-based style. To keep in line with the three computational phases in protein inference, the results obtained from the work of each computational phase consist of the main content of this thesis. They are presented in the form of published or submitted manuscripts. In each chapter, a brief introduction is included to describe the connection of the manuscript to the context of the thesis. Also, a general discussion of the links of each manuscript to the thesis as a whole is also provided in Chapter 6. The paper manuscripts have been modified in format to be consistent with the rest of the thesis. The remaining thesis is structured as follows: Chapter 2 introduces the background of mass spectrometry in proteomics, and a comprehensive review of protein inference is also included. Chapter 3 presents a novel method to determine the charge states of MS/MS spectra from low-resolution collision induced dissociation (CID). This is one of the important operations in processing MS/MS data. It can save a lot of time and resources in performing database searching of peptide identification. Chapter 4 proposes a method based on logistic regression (LR) to compute the probability of peptide identification. The results of this work are used as input into protein inference. Based on the work introduced in Chapter 4, a unified framework and an iterative method are developed in Chapter 5 to infer proteins and identify peptides simultaneously. Protein inference and peptide identification are combined together by adding a feedback from protein inference to peptide identification. Finally, the main conclusions and contributions of this thesis, and some recommendations for future work are summarized in Chapter 6. In addition, a general discussion is given to summarize the relationship of each manuscript to the thesis. The full list of publications is included in Appendix A, and the copyright permissions of included manuscripts are in Appendix B.

References

[1] A. Pandey and M. Mann, "Proteomics to study genes and genomics," *Nature*, 405: 837-846, 2000.

[2] A. I. Nesvizhskii and R. Aebersold, "Interpretation of shotgun proteomic data: the protein inference problem," *Mol. Cell. Proteomics*, 4(10): 1419-1440, 2005.

[3] M. P. Washburn, "Large-scale analysis of the yeast proteome by multidimentional protein idetification technology," *Nat. Biotechnol.*, 19: 242-247, 2001.

[4] R. Aebersold and D. R. Goodlett, "Mass spectrometry in proteomics", *Chem. Rev.*, 101: 269-295, 2001.

[5] E. Kolker and R. Higdon and J. M. Hogan, "Protein idetification and expression analysis using mass spectrometry", *Trends Microbiol.*, 145: 229-235, 2006.

[6] D. L. Tabb, H. McDonald, J. R. Yates III, "DTASelect and contrast: tools for assembling and comparing protein identifications from shotgun proteomics", *J. Proteome Res.*, 1(1): 21-26, 2002.

[7] Z. He, C. Yang, and W. Yu, "A partial set covering model for protein mixture identification using mass spectrometry data," *IEEE Trans. Comput. Biol. Bioinf.*, 8(2): 368-380, 2011.

[8] T. Huang, J. Wang, W. Yu, and Z. He, "Protein inference: a review," *Brief. Bioinform.*, 2012, 13:586-614.

[9] I. J. Good, "Explicativity: a mathematical theory of explanation with statistical applications," *Proc. R. Soc.*, 354: 303-330, 1997. London.

[10] A. I. Nesvizhskii, A. Keller, E. Kolker and R. Aebersold, "A statistical model for identifying proteins by tandem mass spectrometry," *Anal. Chem.*, 75: 4646-4658, 2003.

[11] B. Zhang, M. C. Chambers and D. L. Tabb, "Proteomic parsimony through bipartite graph analysis improves accuracy and transparancy," *J. Proteome Res.*, 6: 3549-3557, 2007.

[12] T. S. Price, M. B. Lucitt, W. Wu, D. J. Austin, A. Pizarro, A. K. Yocum, I. A. Blair, G. A. FitzGerald and T. Grosser, "EBP: a program for protein identification using multiple tandem mass spectrometry datasets," *Mol. Cell. Proteomics*, 6: 527-536, 2007.

[13] B. Lu, A. Motoyama, C. Ruse, J. Venable, and J. R. Yates III. "Improving protein identification sensitivity by combining MS and MS/MS information for shotgun proteomics using LTQ-Orbitrap high mass accuracy data". *Anal. Chem.*, 80(6):2018C25, 2008.

[14] J. Li, L.J. Zimmerman, B.H. Park, D.L. Tabb, D.C. Liebler and B. Zhang, "Network-assisted protein identification and data interpretation in shotgun proteomics", *Mol. Syst. Biol.* 5:303, 2009.

[15] S.R. Ramakrishnan, C. Vogel, T. Kwon, L.O. Penalva, E.M. Marcotte, and D.P. Miranker, "Mining gene functional networks to improve mass-spectrometry based protein identification", *Bioinformatics*, 25(22):2955C61, 2009.

[16] S.R. Ramakrishnan, C. Vogel, J.T. Prince, Z. Li, L.O. Penalva, M. Myers, E.M. Marcotte, D.P. Miranker and R. Wang, "Integrating shotgun proteomics and mRNA expression data to improve protein identification", *Bioinformatics*, 25(11):1397C403, 2009.

[17] S. Gerster, E. Qeli, C. H. Ahrens and P. Buhlmann, "Protein and gene model inference based on statistical modeling in k-partite graphs," *PNAS*, 107(27): 12101-12106, 2010.

[18] M. Spivak, D. Tomazela, J. Weston, M. J. MacCoss, and W. S. Noble, "Direct maximization of protein identifications from tandem mass spectra," *Mol. Cell. Proteomics*, 2012.

[19] B. Kuster, M. Schirle, P. Mallick and R. Aebersold, "Scoring proteomes with proteotypic peptide probes," *Nat. Rev. Mol. Cell Biol.*, 6: 577-583, 2005.

[20] R. Craig, J. P. Cortens and R. C. Beavis, "The use of proteotypic peptide libraries for protein identification," *Rapid Commun. Mass Spectrom.*, 19: 1844-1850, 2005.

[21] H. Tang, R. J. Arnold, P. Alves, Z. Xun, D. E. Clemmer, M. V. Novotny, J. P. Reilly and P. Radivojac, "A computational approach toward label-free protein quantification using predicted peptide detectability," *Bioinformatics*, 22: e481-e488, 2006.

[22] J. A. Taylor and R. S. Johnson, "Sequence database searches via de novo peptide sequencing by tandem mass spectrometry," *Rapid Commun. Mass Spectrom.*, 11: 1067-1075, 1997.

[23] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby and G. Lajoie, "PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry," *Rapid Commun. Mass Spectrom.*, 17: 2337-2342, 2003.

[24] L. Mo, D. Dutta, Y. Wan and T. Chen, "MSNovo: A dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry," *Anal. Chem.*, 79: 4870-4878, 2007.

[25] D. N. Perkins, D. J. C. Pappin, D. M. Creasy and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, 20: 3551-3567, 1999.

[26] J. K. Eng, A. L. McCormack and J. R. Yates III, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *J. Am. Soc. Mass Spectrom.*, 5: 976-989, 1994.

[27] R. Craig and R. C. Beavis. "TANDEM: matching proteins with tandem mass spectra," *Bioinformatics*, 20: 1466-1467, 2004.

[28] A. Keller, A. I. Nesvizhskii, E. Kolker and R. Aebersold,"Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search," *Anal. Chem.*, 74: 5383-5392, 2002.

[29] L. Kall, J. D. Storey, M. J. MacCoss, and W. S. Noble, "Assigning significance to peptides identified by tandem mass spectrometry using decoy databases," *J. Proteome Res.*, 7: 29-34, 2008.

[30] H. Choi and A. I. Nesvizhskii, "Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling," *J. Proteome Res.*, 7: 286-292, 2008.

[31] H. Choi, D. Ghosh, and A. I. Nesvizhskii, "Assigning significance to peptides identified by tandem mass spectrometry using decoy databases," *J. Proteome Res.*, 7: 47-50, 2008.

[32] H. Choi, and A. I. Nesvizhskii, "Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics", *J. Proteome Res.*, 7(1):254C65, 2008.

[33] J. Shi, W. Lin, and F.-X. Wu, "Statistical analysis of Mascot peptide identification with active logistic regression," *iCBBE*, 2010.

[34] J. Shi, and F. X. Wu, "Assigning probabilities to Mascot peptide identification using logistic regression," Advances in Experimental Medicine and Biology, 1, Volume 680, Advances in Computational Biology, Part 3: 229-236.

[35] J. Shi, B. Chen and F. X. Wu, "Improve accuracy of peptide identification with consistency between peptides," *IEEE BIBM*, 191-196, 2011.

[36] T.D. Veenstra, T.P. Conrads, and H.J. Issag. "Commentary: what to do with one-hit wonders?", *Electrophoresis*, 25: 1278-1279, 2004.

[37] J. Klimek, J. S. Eddes, L. Hohmann, J. Jackson, A. Peterson, S. Letarte, P. R. Gafken, J. E. Katz, P. Mallick, H. Lee, A. Schmidt, R. Ossola, J. K. Eng, R. Aebersold and D. B. Martin, "The standard protein mi database: a diverse data set to assist in the production of improved peptide and protein identification software tools," *J. Proteome Res.*, 7: 96-103, 2008.

[38] A. Keller, S. Purvine, A. I. Nesvizhskii, S. Stolyar, D. R. Goodlett and E. Kolker, "Experimental protein mixture for validating tandem mass spectral analysis," *OMICS*, 6(2): 207-212, 2002.

[39] O. Schulz-Trieglaff, N. Pfeifer, C. Gropl, O. Kohlbacher, and K. Reinert, "LC-MSsim - a simulation software for liquid chromatography mass spectrometry data", *BMC Bioinformatics*, 9:423, 2008.

[40] A.M. Zou, J. Shi, J. Ding and F. X. Wu, "Charge state determination of peptide tandem mass spectra using support vector machine (SVM)," *IEEE Trans. Inf. Technol. Biomed.*, 14(3): 552-558, 2010.

[41] J. Shi, and F. X. Wu, "Peptide charge state determination of tandem mass spectra from low-resolution collision induced dissociation," *BMC Proteome Science*, 9(Suppl 1):S3, 2011.

[42] J. Ding, J. Shi, and F. X. Wu, "Quality assessment of tandem mass spectra by using a weighted k-means," *Clinical Proteomics*, 5(1): 15-22, 2009.

[43] J. Ding, J. Shi, and F. X. Wu, "SVM-RFE based feature selection for tandem mass spectrum quality assessment," *Int. J. Data Min. Bioinform.*, 5(1): 73-88, 2011.

[44] J. Shi, B. Chen and F. X. Wu, "Unifying protein inference and peptide identification with feedback to update consistency between peptides," *Proteomics*, 2012, accepted.

[45] P. Alves, R. J. Arnold, M. V. Novotny, P. Radivojac, J. P. Reilly and H. Tang, "Advancement in protein inference from shotgun proteomics using peptide detectability," *Pac. Symp. Biocomput.*, 12: 409-420, 2007.

[46] Y. F. Li, R. J. Arnold, Y. Li, P. Radivojac, Q. Sheng and H. Tang, "A Bayesian approach to protein inference problem in shotgun proteomics," *J. Comput. Biol.*, 16:1183-1193, 2009.

[47] J. Shi, and F. X. Wu, "A feedback framework for protein inference with peptides identified from tandem mass spectra," *Proteome Science*, 2012, accepted.

[48] C. Shen, Z. Wang, G. Shankar, X. Zhang and L. Li, "A hierarchical statistical model to assess the

confidence of peptides and proteins inferred from tandem mass spectrometry," *Bioinformatics*, 24: 202-208, 2007.

[49] Q. Li, M. MacCoss and M. Stephens, "A nested mixture model for protein identification using mass spectrometry," *Ann. Appl. Stat.*, 4(2): 962-987, 2010.

[50] Z. He, H. Zhao and W. Yu, "Score regularization for peptide identification," *BMC Bioinfomatics*, 12(Suppl):S2, 2011.

Chapter 2

Mass spectrometry

Mass spectrometry (MS) is an important technique used in the study of proteomics. Protein inference and peptide identification are mainly achieved by analyzing and interpreting MS data. This chapter will introduce the process of acquiring MS data, followed by an introduction of preprocessing and interpreting MS data. In addition, though protein inference and peptide identification are computational steps (i.e. bioinformatics) in proteomics, it is helpful for researchers who focus on the computation to know the whole workflow of a proteomics project. When we can see the position of our work clearly in the big picture, we would know better the importance and the direction of our work. Hence, this chapter will start with the introduction of general workflows in proteomics, which include targeted and non-targeted ones. After that, considerations and processing that are needed to be taken before mass spectrometry, during mass spectrometry and after mass spectrometry are introduced.

2.1 Proteomics workflow

Proteomics is the large-scale study of proteins, which includes the identification and characterization of proteins, and the study of interactions between proteins [1]. The core instrument in proteomics is a mass spectrometer. According to the entities introduced into a mass spectrometer, protein inference and characterization by MS can be classified into top-down and bottom-up proteomics [2]. In the top-down approach,

intact protein ions or large protein fragments are subjected to gas-phase fragmentation for MS analysis. In the bottom-up approach, purified proteins or protein mixtures are enzymatically digested into peptides, and the resulting peptides are subjected to MS analysis. The top-down approach is relatively young, and its application is limited by the determination of multiply charged product ion masses. Whereas the bottom-up approach is mature and has been widely used in proteomics labs. In this chapter, we will focus on bottom-up proteomics, which can be further divided into targeted and non-targeted proteomics. The workflows of these two approaches are introduced.

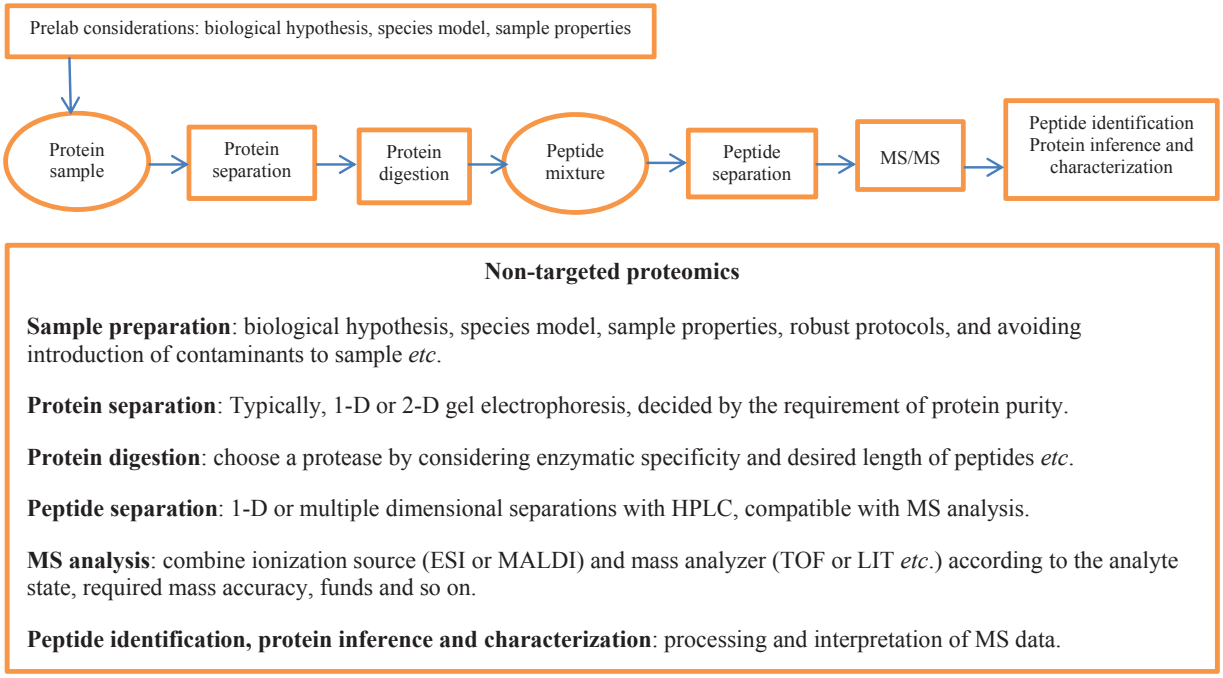


Figure 2.1: Basic steps of MS analysis for non-targeted proteomics with notes for each step.

2.1.1 Non-targeted proteomics

Proteins of interest are not known in non-targeted proteomics. The inference and characterization of such proteins in a sample relies on the MS analysis. The basic steps of MS analysis are shown in Figure 2.1. Under this workflow, proteomics experiments tend to generate a very high redundancy of tandem mass

spectrometry (MS/MS) data, while having a very limited sensitivity. The reasons arise from both the nature of biological samples and the properties of the adopted mass spectrometers. The nature of biological samples includes:

- The majority of proteins in a sample are of low abundance, and they are hard to detect and identify.

- Only a very small portion of sample proteins are of high concentration, and they dominate the generation of MS/MS data. Typically, these proteins are also not the ones of interest.

- The existence of homologous proteins, especially in eukaryotes, produces a high number of degenerate peptides, which brings much ambiguity in the protein inference step.

- Most of the proteins in a cell will go through certain kinds of modifications during the metabolism of cell growth; these single or multiple modifications complicate the identification of peptides, and they will be lost if modifications are not considered in peptide identification.

- The physicochemical properties of proteins and peptides cause differences in the tryptic digestion of proteins, the ionization ability and the fragmentation sites of peptides, and these differences make some peptides more detectable than others.

And the properties of mass spectrometers include:

- The inevitable introduction of electric and chemical noise into the mass spectra;

- In each duty cycle of a mass spectrometer, high intensity peptides will be selected, and some of them will be repeatedly selected. As a result, redundant mass spectra will be produced for these peptides. While peptides with low abundance may never get a chance to be selected to be analyzed. This suppresses the detection of low-abundance peptides.

In non-targeted proteomics, all proteins in a sample get the same chance to be analyzed, because proteins of interest are unknown. Although there have been incremental improvements to this workflow, its intrinsic weaknesses will need more efforts to be overcome. Currently, to detect the low-abundance proteins and to better understand the change of proteins in the development of diseases, a different proteomic workflow can be employed.

2.1.2 Targeted proteomics

Different from non-targeted proteomics, targeted proteomics first finds the proteins of interest by analyzing the pathological or physiological models. The workflow is given in Figure 2.2. Under this workflow, physiological models are analyzed with the most suitable proteomics technologies, and the changes or differences in proteins between experimental conditions and controls are revealed [3]. This approach allows for the optimal discovery of the changes which define the model system. When performing targeted proteomics, there are two basic problems we need to consider, which are listed as follows.

- How to determine the proteins of interest?

  The determination of these proteins needs a deep understanding of the studied models and the hypotheses. Usually, these proteins are determined by biological experts, or by gleaning from literature if the models have been studied under the same or similar conditions.

- How to use these proteins to improve proteomic results?

  Given proteins of interest, techniques, such as protein chips, can be used to select these proteins from the sample. In this way, the concentration and purity of these proteins of interest can be increased. Another possible use is to create an inclusion list in the real-time generation of MS/MS data in mass spectrometers. This can increase the possibility of generating high-quality MS/MS data for these proteins. But this inclusion list only works when the abundance of proteins is above the detection limit of a mass spectrometer.

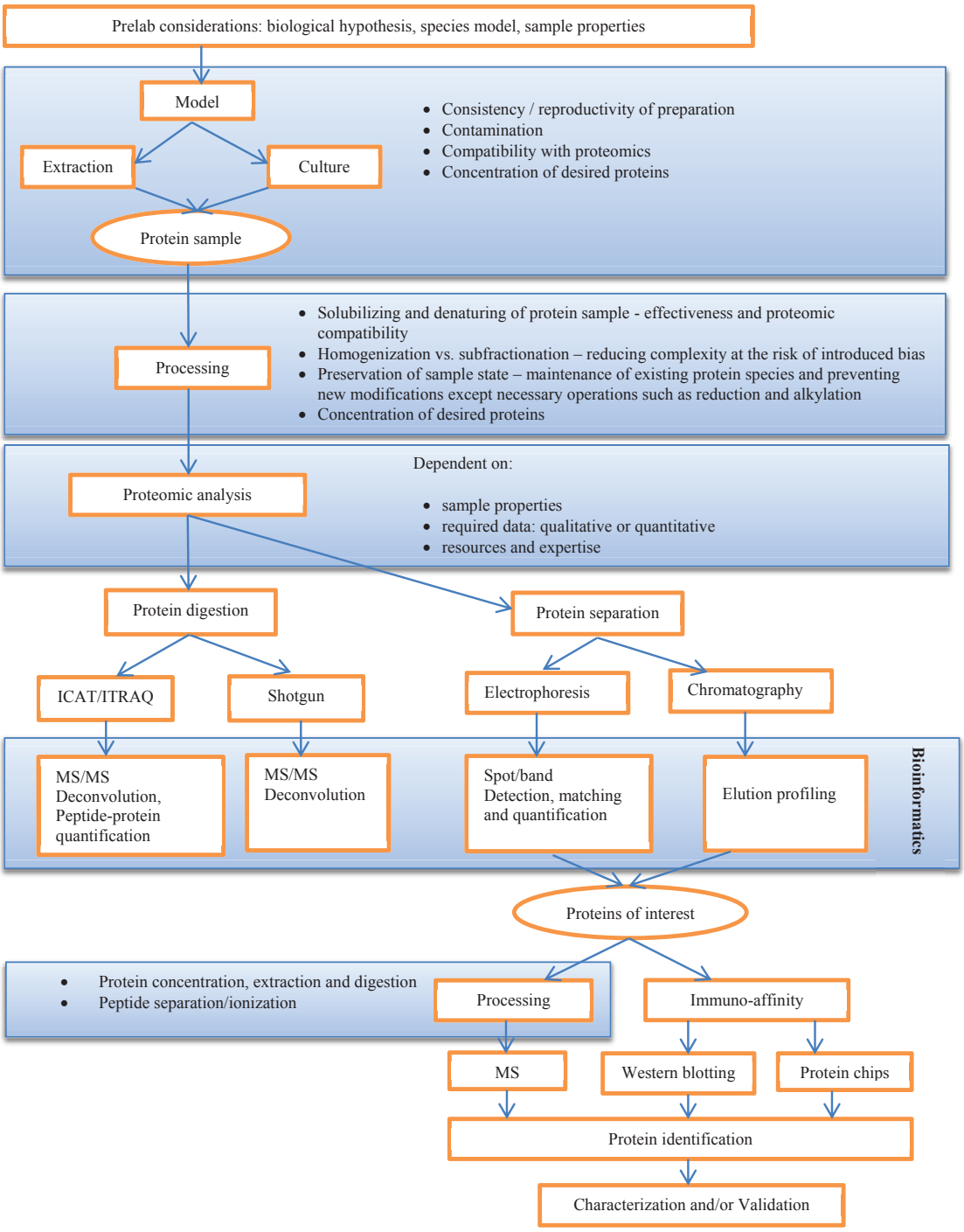


Figure 2.2: An overview of potential steps in proteomic analysis. This figure is adapted from [3]. The differences of targeted and non-targeted proteomics mainly arise from whether the proteins of interest are known or not. If known, then they can be identified with other approaches like antibody-affinity other than MS analysis.

2.2 Before lab experiments

A proteomics project can be very big and needs the collaboration of many researchers. Although there have been projects that are designed to provide reference data for testing algorithms and models proposed for processing proteomics data [4, 5], more general proteomic projects are initiated to provide insights into the real biological models, so that pathologists or physiologists can verify the underlying mechanisms of diseases and disease processes [3]. Thus, before we begin a proteomics project, we first need to understand what we are going to analyze and what we are expecting from the project. This section will outline some considerations we need to take before going into the experimental stage.

2.2.1 Build a hypothesis

Building a hypothesis is one of the most critical steps in a proteomics project, because it will determine the follow-up design of the whole experiment. As pointed out in [3], the key to increasing success in any proteomics experiment is to have a comprehensive understanding of the physiological model or disease process being studied. From the understanding, a hypothesis can be formed, and it will drive the selection of the particular proteomics/analytical approach. For example, the experimental design will differ if the hypothesis involves the change of abundant proteins *versus* the change of low-abundance proteins determining the cause of a disease. The latter one will need special care to improve the protein concentration in the sample preparation step, or to remove some of the high level proteins with protein depletion kits.

2.2.2 Choose a model

The study of proteomics relies on the integration of mass spectrometry and protein databases, and these databases are derived from genomics. Thus, the lack of genomic information of a species can greatly limit

the success of a proteomic project, especially when it comes to the phase of protein inference. It is therefore important for researchers to choose a model of which the genome is completely sequenced.

2.2.3 Consider sample attributes

Considerations should be given to the following properties of a sample: protein concentration, dynamic range of protein concentration, protein solubility in the solvent, and the copy number of protein classes in which one is attempting to assess the changes. It is a rule of thumb to expect reasonable success in identifying the protein of interest if it can be visualized by Coomassie blue staining [3, 6].

2.3 Lab experiments

As shown in Figure 2.2, there are many kinds of operations involved in a proteomics lab experiment. This section will introduce some basic operations that are necessary for MS analysis, which are categorized as operations before MS, in MS and after MS.

2.3.1 Before mass spectrometry

Sample preparation

Sample preparation is a critical step in proteomics experiments. It is important to minimize variations of sample preparation by strictly adhering to robust experimental protocols [7]. An early checking of sample quality and replacing the poor quality samples with high quality ones are preferred. For example, we can check the concentration of proteins to make sure that they are abundant enough to be detected in mass spectrometers. This can prevent a lot of quality problems when it comes to subsequent data analysis. In

addition, it can increase the number of peptides to be identified and therefore improve the coverage of protein inference.

Protein separation

Protein separation is necessary when the complexity of the sample is not suitable for MS analysis for protein inference. The most commonly used technique for protein separation is gel electrophoresis, which separates proteins according to molecular mass or isoelectric point. Most separation methods can be described by [8]: (1) the substance through which the molecules migrate; (2) the external force that causes the molecules to migrate; (3) the preprocessing of molecules that enables them to migrate through the substance.

Based on molecular mass, protein separation is performed with SDS-PAGE. SDS is short for sodium dodecyl sulfate and PAGE is PolyAcrylamide Gel Electrophoresis. Strictly speaking, this technique separates molecules according to their size, which are usually proportional to their mass. In SDS-PAGE, the substance is polyacrylamide, which is formed into a porous gel with many small tunnels. This network of tunnels will impede the movement of large molecules while small ones can move much more readily through it. The force applied to move the molecules through the gel is supplied by an electric field. To use PAGE for separation, proteins should be prepared to have a size proportional to their mass. In order to make sure they are separated by their sizes in the electric field, the charges carried by proteins should be proportional to their mass as well, and these charges should be all positive or all negative. Therefore, the necessary preprocessing of proteins includes: (1) denaturing proteins into a linear form to have a mass-proportional size; (2) making them carry mass-proportional charges. These are achieved by treating proteins with SDS, which is a detergent molecule with a long hydrophobic tail and a negatively charged head. SDS can attach to protein sequences to denature them into a linear form, and impart negative charges to them roughly proportional to their sizes. The resulting proteins are loaded onto polyacrylamide gel. The gel is then placed in an electric field, which moves the negatively charged proteins towards the positive electrode with velocities inversely

proportional to their sizes. Finally, proteins are separated by their sizes in the gel.

Based on the isoelectric point (pI), proteins are separated by the technique called isoelectric focusing (IEF). A pH gradient is used for separation. First, proteins enter into the gradient by absorption from a buffer with the sample. Next, an electric field is applied across the gradient. This makes proteins initially move towards the electrode with the opposite charge. As a protein reaches the point that is equal to its pI, its net charge becomes zero, and its migration will stop. Thus, when all proteins reach their respective pI point, they are separated in the gradient. SDS-PAGE and IEF can be combined to perform two-dimensional separation, since they separate proteins on orthogonal attributes. Proteins are separated on pI in the first dimension, and on mass in the second dimension. IEF is performed first because SDS can attach to proteins and make them all carry mass-proportional negative charges, and this makes it inappropriate for proteins to be separated by their pI.

Protein digestion

One important step in MS-based protein inference is to cleave proteins into peptides. The most often used method for protein cleavage is enzymatic cleavage. The enzymes that perform protein cleavage are called proteases. Two basic rules in choosing a protease for protein digestion are:

- The protease should cleave proteins in a consistent and predicable way, that is, it should cleave proteins at some specific sites. This provides some guidance in choosing a protease for the sample under study, and also helps in peptide identification;

- The protease should cleave proteins into peptides of lengths suitable for MS analysis. Mass spectrometers are usually set by users to have a limited mass range, and only peptides with masses (more accurately, mass-to-charge m/z) falling into this range have a chance to be detected. Peptides which are too long or too short will fall out of the range and cannot be detected. Besides, the number of

peptides that share a specific mass increases with the decreasing mass [8]. Thus, short peptides (less than six amino acids) are usually very difficult to discriminate and not suitable for the identification of peptides. Also, because the chemical background is more intense in the low m/z region, therefore peptides of m/z less than 300 are not usually examined.

Trypsin is the most commonly used enzyme for protein digestion. It can produce peptides most suitable for MS analysis. Trypsin cleaves proteins after arginine (R) and lysine (K), except followed by proline (P). Following are some attributes of trypsin [8]: (1) High specificity, with rare missed and unexpected cleavages; (2) Peptides produced are of suitable lengths. R and K appear with an average distance of approximate 11 residues, and with a small probability of being followed by P; and (3) It is easy to be obtained and purified, and is applicable in most experimental settings. Trypsin can be used to cleave proteins in solution, gels, or even can be adsorbed onto surfaces.

Trypsin is suitable for positively charged MS analysis. Peptides need charged to be detected by mass spectrometers. Since R and K are basic residues, peptides produced by trypsin with R or K on the C-terminal have the ability to retain protons. However, trypsin may not be suitable for digesting proteins which are highly basic or highly acidic. Highly basic proteins may contain too many R and K, which will be cleaved into many too small peptides by trypsin. On the other hand, highly acidic proteins will contain many glutamic (E) and aspartic (D) acids while few R and K, which will be cleaved into a few too long peptides by trypsin. In this case, alternative proteases will be needed.

Peptide separation

After protein digestion, we get peptide mixture solutions. To reduce the complexity of MS/MS data, peptide mixtures are subject to further separation with chromatography. Chromatography includes a family of techniques that are used to separate a mixture into its individual components. The most often used chromatography in peptide separation is liquid chromatography (LC). LC uses a liquid as the mobile phase and

a porus solid as the stationary phase. High pressure (HP) is usually applied to change the flow rate of the mobile phase and improve the separation efficiency. HPLC is often classified according to the principle of separation: hydrophobicity, charge, affinity to special functional groups or component size. In proteomics, two often used HPLC are reverse phase HPLC (RP-HPLC), which separates peptides based on hydrophobicity, and strong cation exchange HPLC (SCX-HPLC), which separates peptides based on charges.

The stationary phase in RP-HPLC is the modification of carbon chains with different lengths (like C_4, C_8, C_{18}). The longer the carbon chain, the stronger the hydrophobic interaction between peptides and the stationary phase. Two solutions (A and B) are used as the mobile phase. Solution A is usually water with a small amount of organic acid, in which peptide sample is injected into the column. After being forced through the column, peptides attach to the carbon chains and stay on the stationary phase. To detach peptides, solution B that is mainly an organic solvent is gradually mixed into solution A. With the increase of the concentration of organic solvent, less hydrophobic peptides will detach and move along with the mobile phase to be eluted. More hydrophobic peptides will detach at higher percentage of organic solvent. This change in solvent strength over time is called gradient. When the gradient reaches a certain percentage of solvent, all peptides are usually eluted from the stationary phase.

SCX-HPLC is an ion exchange chromatography, which uses the principle that opposite charges attract each other. Peptides are zwitterionic molecules and their net charge depends on the pH of the solution and their pI. When the pI is above the pH, the peptide is positively charged; otherwise, it is negatively charged. The stationary phase in SCX-HPLC is often a surface modified with sulfonic acid groups, which becomes negatively charged at a pH above $2 \sim 3$. The peptide sample is injected into the column at a low pH solution (often $3 \sim 3.5$). The pI of most peptides is $4 \sim 7$ [8]. Thus, the peptides are positively charged, and they will interact with the negatively charged stationary phase. The more positive charges a peptide carries, the stronger it interacts with the stationary phase. Similar to RP-HPLC, the other solution B is mixed into the pH solution to elute peptides. Usually, solution B contains salts which carry both positive and negative charges. By gradually increasing the salt concentration, these ions will compete with the positive charges on

peptides and negative charges on the stationary phase. The peptides with the weakest binding will detach first and start moving with the mobile phase. While the peptides with stronger binding only detach at a higher salt concentration. Different from RP-HPLC, the eluate from SCX-HPLC cannot be directly infused into a mass spectrometer, because the mobile phase in SCX contains salts which will interrupt the subsequent data acquisition.

2.3.2 Mass spectrometry

Mass spectrometry is an analytical technique which is widely used in the measurement of molecular masses, by recording m/z values of the charged molecules. It is the method of choice for peptide and protein identification today [9–11]. Schematically, a mass spectrometer includes three parts: ionization source, mass analyzer and detector. Although different mass spectrometers have different properties, they have the same underlying principle.

Principle of mass spectrometry

All mass spectrometers use electric or electromagnetic fields to control the movement of charged particles and separate them accordingly. Hence, the molecules to be analyzed need to be ionized before their masses can be measured. The ionized molecules are sent into a mass analyzer, in which they are separated based on their m/z values. The separated molecules then hit a detector, and a mass spectrum is constructed by a connected computer. A mass spectrum is typically shown as a diagram, with m/z on the horizontal axis and the intensity of the signal for each molecule along the vertical axis. Since the analyzer works on m/z but not on the mass directly, the charge of a molecule must be known to determine the mass.

Ionization source

There are many ionization sources, however, the two most commonly used ionization sources in proteomics laboratories are, electrospray ionization (ESI) and matrix assisted laser desorption/ionization (MALDI). MALDI is mainly used in peptide mass fingerprinting as it predominantly yields singly charged ions. It is more tolerant to salts and contaminants compared with ESI, and is usually used for samples with a small number of proteins. ESI typically produces multiply charged ions and is applied in MS/MS analysis. It is used for more complicated samples, because it is readily coupled to the LC system. ESI forms ions from solutions whereas MALDI requires to spot the analytes on a plate with a suitable matrix and laser ionization.

Mass analyzer

The basic operation of a mass analyzer is to separate peptides and measure their m/z values. While an analyzer for MS/MS performs two tasks: one for selecting the m/z of interest, and the other one for measuring the m/z values of the fragment ions. These tasks can be performed in two analyzers, called in-space analyzers, or in one analyzer at different times, called an in-time analyzer. In order to allow both MS and MS/MS analysis to be performed on one instrument, analyzers commonly can function in two scanning modes. Take an example of an ion trap analyzer, in full-scan mode, all peptide ions from the ionization source are analyzed and retained, allowing the recording of a mass spectrum. In MS/MS mode, the analyzer only retains ions that fall within the specified m/z range, ejecting or cutting off any ions that fall out of this range. After the selection, the retained ions are fragmented into fragment ions. Then, the fragment ions are analyzed in the full-scan mode, which produces a tandem mass spectrum. Mass spectrometers are configured to continuously switch between these two modes, and automatically record MS and MS/MS data. The mechanism used to measure the m/z of ions depends on the type of an analyzer. These can be based on the time of flight of the ions, their movement in magnetic or electromagnetic fields and so on.

2.3.3 After mass spectrometry

After tandem mass spectra are obtained, the subsequent work is to process, analyze and interpret the data. This work includes raw data processing, peak list processing, peptide identification, protein inference, protein quantification and characterization. Peptide identification and protein inference will be introduced in the next two sections. Here we will elaborate a little bit on raw data and peak list processing, which are necessary steps to improve the peptide and protein identification results.

Raw data processing

Raw MS data are continuous with peaks and valleys, which are not convenient for data analysis and interpretation. They are thus converted into peak lists, usually by the in-house software specific to each mass spectrometer. However, some basic operations and considerations are the same.

First, one has to consider the noise in a spectrum. The quality of a spectrum is strongly influenced by the amount of noise *versus* the amount of signal peaks (corresponding to peptides in MS or peptide fragments in MS/MS) in the spectrum. There are two main types of noise:

- Chemical noise: the sources of chemical noise may be the contaminants introduced during sample handling, like the detergents not removed from the sample and polymers from plastic tubes, or proteins unintentionally brought into sample, such as keratin from the skin or hair.

- Electric noise: Electronic noise comes from the electronic disturbances, and happens with random fluctuations between the chemical noise.

A form of noise occurs as the baseline of a spectrum, derived predominantly from chemical noise [8]. The baseline is an offset of the intensities of masses, and should be subtracted from the measured intensities. It is usually dependent of m/z values, such that it is highest at low m/z values, and decays toward higher m/z

values. The simplest method to remove the baseline is to subtract the lowest point in the spectrum. Note that the baseline varies from spectrum to spectrum, so it should be treated individually to each spectrum.

Secondly, one needs to detect peaks from noise and pick them to construct the peak list. The goal is to represent each peak with exactly one data point. One way is to identify the apex of a peak, and the intensity at this point is compared to the surrounded noise level to determine the start and end point of this peak. Another way is to use the valley on the both sides of the apex to determine the start and end point of the peak. The area of the peak is then calculated, and is compared to a threshold to decide whether this peak is a signal or noise. After a signal peak is detected, one needs to compute its m/z value and intensity in the derived spectrum. The intensity is proportional to the area under the peak, and the value at the centroid of the peak is usually used to calculate the ion's actual m/z value.

Finally, the spectra derived have to be calibrated to achieve the accuracy required for a database search, because there usually exists a mass shift in the spectra. The most commonly used technique is internal calibration. This is achieved by adding known standards to the sample, and determining the exact m/z values of certain peaks from the standards in the spectrum. Another way is to find peaks from the autolysis products of the used protease. Measurement deviations, which are observed for the known peaks, are then used to compute a function to calibrate the masses of other peaks.

Peak list processing

Peak lists derived from raw MS data need to be further processed such that they are more appropriate for peptide and protein identification.

- Monoisotoping and deisotoping

  This step reduces a cluster of isotopic peaks to a single peak, with intensity equal to the sum of the isotope intensities. Monoisotoping reduces isotopic cluster to the peak with the lowest m/z in the

cluster. Deisotoping reduces the isotopic cluster to a centroid peak, with m/z value determined from the intensities of the individual isotopes. The centroid m/z value is obtained with the average masses of the atoms in the peptide used for calculating its mass.

- Denoising

  Although the initial peak construction eliminates some noise, the derived discrete spectra are still noise contaminated. To achieve good matches in a database search, the spectra should be further denoised with more complex methods by considering their properties [12, 13]. In addition, some known contaminants, such as peptides from keratin or autolytic peptides from the protease should also be filtered.

In addition to the above processing which actually changes the peak lists in MS/MS, other analysis can also be performed to facilitate peptide identification, which includes quality assessment and charge state determination of MS/MS data. Machine learning methods such as support vector machine, and Gaussian mixture model are applied to such an analysis by using significant features to describe MS/MS [14–16]. This analysis can improve the chance of identifying true peptides in a database search, and also significantly save time in the searching step.

2.4 Peptide identification

Peptide identification is the first computational step in proteomics. Its accuracy is critical to the success of the subsequent protein inference [17]. Database searching and de novo sequencing are the two main methods for peptide identification, which are introduced in this section. Also, the target-decoy database used to evaluate the identification results is also introduced.

2.4.1 Database searching

Database searching is the dominant method for peptide identification in proteomics. The procedure of this method is: First, database proteins are cleaved to produce peptides in terms of enzyme specificities; Second, theoretical MS/MS of these peptides are generated; Third, a scoring system is used to measure the similarity between the experimental MS/MS and the theoretical MS/MS, i.e., performing peptide-spectrum-match (PSM); Finally, the peptide with the highest score is usually reported to be identified. Many search engines have been developed for peptide identification [18–22], and the main difference between them lies in the scoring system. This leads to the situation that one query spectrum will have two different sets of scores after being searched with, for instance, Mascot [18] and Sequest [19]. As such, it is hard to compare the identification results with these scores. In addition, peptide identification is only an intermediate step in proteomics. It lays the foundation of protein inference. To facilitate the comparison between the identification results and the subsequent protein inference, statistical analysis of peptide identification results is usually performed [23–25].

Database searching has several obvious advantages in peptide identification. First, it is very simple and natural in practice. Once a protein sequence database is available, peptide identification by database searching would be very simple to implement by well-developed programs [18–22]. Second, a database always has a limited searching space while the de novo sequencing does not. In addition, databases have been growing very fast in size in recent years and this means their completeness is also growing. The completeness of a database is critical, because we can never find peptides that are not in a database. Actually, this leads to a conflict between the need of a small searching space, which can reduce computational effort, and the need of database completeness, which can increase correct identifications. Fortunately, databases with both satisfactory completeness and relatively small searching space can be formed with the observation of proteotypic peptides [26–28].

Although a database search is an effective method for peptide identification, there are also drawbacks of this method. First, it is limited by the used database. On the one hand, the completeness of the database can directly determine the accuracy of the identification results [11, 29]. On the other hand, the increasing size of the database requires more computational effort. This could be a big problem for large-scale and complex sample analysis [29]. Second, the generation of theoretical spectra is not accurate [30]. In silico digestion of proteins is purely based on the "ideal" sites that are cleaved by enzymes (typically trypsin). In contrast, the production of experimental spectra varies a lot due to many factors, such as the uneven probability of being ionized in the competition for protons. Third, a database search cannot identify peptides which are modified in an unexpected way [31]. For example, when new proteins, mutations, post-translational modifications (PTMs) and sequencing errors happen, database searching cannot identify such peptides [32]. Last but not least, statistical analysis of the identification results can be tough because of the variants in the experiment [11, 33]. Now the publication of proteomics data requires or encourages author(s) to provide the software and statistical analysis of their results [34, 35], otherwise their results would not be reliable for other researchers to use.

2.4.2 De novo sequencing

De novo sequencing predicts peptide sequences directly from tandem mass spectra [31]. It has benefits in the situation that effective databases are not available or there exists protein homologies and modifications in the sample under study. Besides, it can be used to validate results from a database search. If de novo sequences explain MS/MS data better than database-derived sequences, then the database-derived sequences are likely to be false positives [36].

Many programs have been developed to implement de novo peptide sequencing. The software package Lutefisk [36, 37] is a typical one which employs graph theory for de novo peptide sequencing. In this approach, the spectrum is first translated into a sequence graph. The nodes in the graph represent peaks in

the spectrum, and two nodes are connected with an edge when their mass difference is close to an amino acid mass. The software then tries to find a path, which connects the N and C termini and connects all the nodes corresponding to y-ions or b-ions. The problem is that it can be complicated and often fail by the absence of ions, which break the path the software tries to find. Later, Ma *et al.* [38, 39] developed a different software package PEAKS, which works directly on spectra without translating them into sequence graphs. In essence, this de novo sequencing can be regarded as trying to identify peptides from the "exhausted" peptide sequence database, which contains all the possible combinations of animo acids. It introduces rewards and penalties for ions to scoring a candidate peptide sequence. Candidate peptide sequences are formed by considering all the possible amino acid combinations. A positive reward is added to a sequence if it can generate a y-ion or b-ion with the mass which is close to a given peak's mass value. Otherwise, if there is no ion's mass close to the given peak's mass value, a negative penalty is added to the sequence. Thus, searching candidate peptides is reduced to finding sequences whose b and y ions can maximize the total rewards at their mass values. Owning to the use of reward and penalty, the absence of ions does not cause as many problems as Lutefisk to PEAKS. Other programs for de novo peptide sequencing can be referred to in [17, 40–44].

2.4.3 Target-decoy database

Studies have shown the lack of consistency in the false-discovery rates (FDRs) of peptide identification when using the thresholds of Sequest [45] and Mascot [46]. This means that we need to validate peptide identification results in an experiment-specific or dataset-specific way. One way to estimate the FDR of peptide identification is to use the decoy database which is formed by reversing or reshuffling protein sequences in the original (target) protein database [45, 47–50]. The assumption is that the occurrences of false discoveries in the target database is equally likely in the decoy database. When searching MS/MS spectra against the decoy database, we are sure that the resultant PSMs are incorrect. They are used as the surrogates for incorrect PSMs obtained by searching against the target database. Then, the FDR is calculated as the number of decoy false positive PSMs over the number of positive target PSMs (including true positive and false

positive PSMs). Generally, there are two ways to use the decoy database. One way is to search the target and decoy database separately and the other is to search a concatenated target-decoy database. Elias *et al.* [47] pointed out that the target-decoy database works better than the separate searches in two databases.

2.5 Protein inference

MS-based protein inference can be performed with one-stage or two-stage MS data. The one based only on MS data is called peptide mass fingerprinting (PMF), and the one based on MS/MS data is by assembling identified peptides to infer proteins. Database searching is the core operation in both methods. In the following, we will describe these two methods with a greater focus on protein inference with MS/MS data.

2.5.1 Peptide mass fingerprinting

PMF identifies proteins by matching observed peptide masses to theoretical peptide masses generated by virtually digesting database proteins. The presumption of PMF is that every protein has a set of unique peptides, and thus masses of these peptides can form its fingerprinting. The performance of PMF heavily relies on the high mass accuracy and precise cleavage of enzymes [31, 51, 52]. The study of PMF was promoted by the advent of a high accuracy mass spectrometer MALDI-TOF in early 1990s. MALDI-TOF predominantly produces singly charged peptides, so it is easy to compute their masses [53]. Usually, PMF has a good performance with 2D gels in which proteins have a high purity, but it can run into troubles when dealing with complex protein mixtures. Additionally, incomplete cleavages of proteins and post-translational modifications can decrease the sensitivity of PMF [31]. Finally, it is challenging work in the future to improve PMF so as to handle more proteins at one time and relax the requirement of sample separation.

2.5.2 Protein inference by assembling peptides

Protein inference based on MS/MS is usually performed in a two-stage way. First, peptides are identified from MS/MS with database searching or de novo sequencing. The identification results are often subject to a statistical analysis [23–25]. At this point, peptide identification is completely finished. Secondly, protein inference is conducted based on the output from peptide identification. This strategy has been widely used in protein inference, and implemented in many programs, which will be briefly introduced in later sections. The shortcoming of this strategy is that there is no message passing between protein inference and peptide identification, which can provide useful information to improve the confidence and increase the number of identified peptides [54, 55]. In turn, the coverage and confidence of protein inference can also be increased [56, 57]. Before introducing the algorithms for assembling peptides to proteins, we first see some common challenges in protein inference, and some possible solutions to these challenges. In particular, an MS/MS intensity-based strategy which was proposed to address the challenge of assigning degenerate peptides is discussed.

Challenges in protein inference

After obtaining statistically reliable peptide identification, protein inference is more than only assembling peptides to proteins in a database. Many challenges exist in this step. First, it is hard to assign degenerate peptides to the protein(s) which truly exist in the sample. Theoretically, the presence of degenerate peptides implies that any protein containing them has a chance to be identified. However, the more realistic chance is that this degenerate peptide only comes from one or the partial proteins but not all those proteins [26, 58]. Second, it is a tough task to develop analysis methods and statistical models for protein inference. Many factors in proteomics experiments influence the inference results. It is expected to integrate all possible factors into one analysis model so as to improve the accuracy of protein inference. For example, different experimental designs can result in different datasets, and a good analysis method should be able to be adapted

to handle these different datasets. Generally, a good model sets parameters which could be adjusted by users so that it works in experiment- and data-specific applications. In addition, it is a challenge to identify low-abundance proteins in a complex sample. It has been shown that low-abundance proteins tend to fail in competing for cleavages in the digestion phase and also often fail in getting protons in the ionization phase [58]. This can shrink their probability to be detected and identified. Finally, a statistical analysis of identification results is as important as the identification itself. It is well known that there exists a high rate of false positives in peptide identification and this rate can be magnified in protein inference.

Useful concepts

Many useful concepts have been proposed to address the challenges mentioned above. In the following, we introduce three of them that are often used and appear frequently in recent research papers.

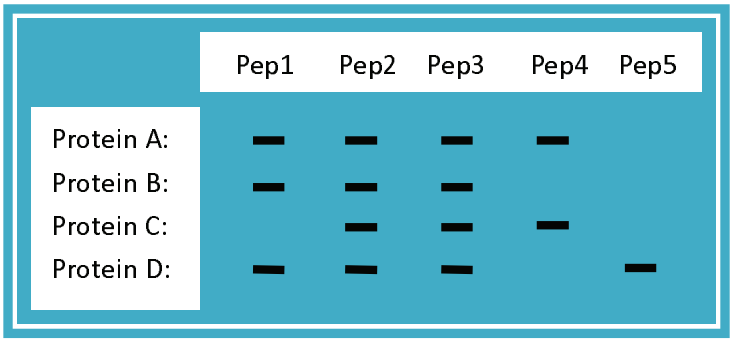


Figure 2.3: Parsimony principle to solve degenerate peptides. Only protein A and protein D would be reported to be inferred because they can explain all the observed peptides. Although protein C and protein D can also explain all the observed peptides, protein A is favored because it can explain more peptides than protein C.

- **Parsimony principle** applies Occam's razor [59] to deal with homologous proteins and degenerate peptides. According to this principle, only the simplest group of proteins which are sufficient to explain all the observed peptides are reported to be inferred [24, 60]. For example, in Figure 2.3, only Protein A and D would be reported because they are enough to explain all the 5 peptides.

- **Proteotypic peptides** are the peptides in a protein that are most likely to be observed by current MS-based proteomics methods [26, 28]. The proteotypicality of a peptide can be predicted according to the peptide's chemico-physical properties [28, 32, 34, 61]. By building proteotypic peptide libraries, protein inference can be based on the identification of proteotypic peptides. Because proteotypic peptides can be identified with a high confidence, the sensitivity of protein identification can consequently be increased.

- **Peptide detectability** is defined as the probability of observing a peptide in a standard sample by a standard proteomics routine [58]. A standard sample is a sample which contains a fixed number of different proteins (peptides), and they are mixed at the same fixed concentration [58]. Further, under this condition, peptide detectability is considered as an intrinsic property of a peptide that is mainly decided by its primary sequence and its parent protein sequence. By this definition of peptide detectability, a degenerate peptide now can be assigned to each of its parent proteins with a corresponding probability assuming that it comes from that protein. Currently, this is a concept that can explain the assignment of degenerate peptides in principle, compared with the use of weights [24], which presumes that degenerate peptides only can come from one protein, or the use of the concept of peptide grouping [29], which assigns two peptide sequences into the same peptide group if their predicted spectra are not distinguishable.

MS/MS intensity-based strategy for assigning degenerate peptides [57]

As discussed before, it is difficult to compute the probabilities of a degenerate peptide belonging to different parent proteins, because the connection between peptides and proteins is lost in proteome experiments. Here we propose an MS/MS intensity-based strategy to assign degenerate peptides to truly present proteins. The idea is that, for a given peptide which is shared by protein Q_1 and Q_2, if the peptide was from Q_1, then its intensity will be closer to the intensity of its siblings in Q_1 than that in Q_2. The intensity of a peptide is

computed with the signal peak intensity in its matched tandem mass spectra.

This MS/MS intensity-based method requires that all peptides in the sample have a similar ability to be ionized and fragmented, and thus have a similar chance to be analyzed by mass spectrometers. However, this is not the case in practice. One way to alleviate the effect of peptide detectability [58] on peptide intensity is that, for each protein with degenerate peptides, we compute the average intensity of peptide siblings, and compare this intensity to the intensity of a degenerate peptide. Some peptides of a protein may have low detectability, but others may not. Thus, averaging the intensity of all peptide siblings can help to reduce the effect of detectability on intensity. An alternative way is to combine peptide detectability into the computation of peptide intensity, if the computation of detectability is accurate enough. The intensity of a peptide P_i is computed as the sum of the signal peak intensity in all its matched tandem mass spectra, which is given by

$$I_i = \sum_{j=1}^{N_s} S_{p_j}, \tag{2.1}$$

where I_i is the peptide intensity and N_s is the number of tandem mass spectra matched to the peptide. The S_{p_j} is the preliminary score in Sequest [19] output for the j^{th} tandem mass spectrum, which is the sum of the intensity of all signal peaks in the spectrum. And it is factored with the ratio between experimental and theoretical peaks that can be derived from the peptide. This factor can eliminate the unfair advantage of longer peptides over short ones. In addition, S_{p_j} is normalized with the maximum value in each whole data set.

As previously mentioned, for a given degenerate peptide, the intensity of its siblings is averaged in order to reduce the effect of peptide detectability on intensity. So the intensity of the siblings of a degenerate peptide P_i is calculated by

$$J_i = \frac{1}{N_i} \sum_{j=1}^{N_i} I_j, \tag{2.2}$$

where J_i is the average intensity of the siblings of peptide P_i, and N_i is the number of its siblings. I_j is the intensity of its j^{th} sibling peptide.

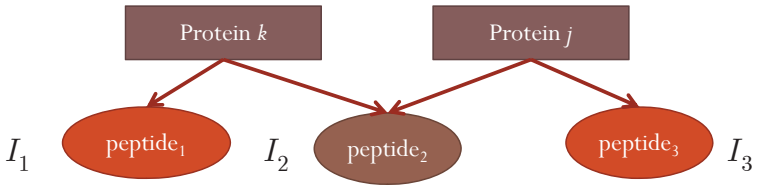


Figure 2.4: A toy example of the assignment of degenerate peptides. The intensity of the three peptides are I_1, I_2, I_3, respectively.

The intensity of a degenerate peptide is contributed to by all of its parent proteins in the sample. This makes the intensity proportion contributed to by each protein sum to unity. A simple example is used to illustrate how to compute these proportions. In Figure 2.4, peptide P_2 is shared by protein Q_k and Q_j. The proportion contributed by protein Q_k to the intensity of peptide P_2 is calculated by

$$P_2^{k'} = \frac{|I_2 - I_3|}{I_2}, \tag{2.3}$$

where $|\cdot|$ is the absolute value operator. Similarly, the proportion contributed by protein Q_j is given by

$$P_2^{j'} = \frac{|I_2 - I_1|}{I_2}. \tag{2.4}$$

Since the proportions contributed by all proteins sum to 1, the previous proportions are normalized,

$$P_2^k = \frac{P_2^{k'}}{P_2^{k'} + P_2^{j'}}, \qquad P_2^j = \frac{P_2^{j'}}{P_2^{k'} + P_2^{j'}}. \tag{2.5}$$

For any given peptide P_i, and its parent protein Q_k, the proportion of the intensity of P_i contributed to by the protein Q_k, denoted by P_i^k, is given as follows,

$$\begin{aligned} P_i^{k'} &= \frac{\left| I_i - \sum_{f \neq k} J_i^f \right|}{I_i} \\ P_i^k &= \frac{P_i^{k'}}{\sum_{\text{All parent Q}_f \text{ of P}_i} P_i^{f'}}, \end{aligned} \tag{2.6}$$

where I_i is the intensity of peptide P_i, and J_i^f is the average intensity of the siblings of peptide P_i from protein Q_f. Here, we take this proportion to represent the probability of peptide P_i belonging to protein Q_k.

It is worth pointing out that although the probabilities of degenerate peptides also sum to 1 as in ProteinProphet [24], it is not required that these shared peptides can only come from one truly present protein in the sample. In the case of ProteinProphet, the weights of a shared peptide will eventually be one of them that becomes close to 1, and the others become close to 0, because it assumes that shared peptides can only come from one truly present protein. This is not true in practical experiments and also misinterprets the real meaning of shared peptides. By removing this assumption, the probability P_i^k allows degenerate peptides to be assigned to multiple proteins in the sample, as long as these proteins have enough evidence to support their existence.

Assembling peptides to a protein list

Protein inference by assembling peptides identified from tandem mass spectra is an important computational step in proteomics, based on which further analysis, such as inference of protein structure and function can be performed. This problem has been systematically discussed in [33, 62, 63]. Existing MS-based methods to address this problem can be divided into two groups. The first group performs protein inference and peptide identification separately [24, 64–66]. First, peptides are identified from tandem mass spectra by de novo sequencing [37, 38, 42] or database searching [18, 19, 21]. Then, proteins are inferred by assembling these identified peptides. The other group combines protein inference with peptide identification, identifying peptides and proteins simultaneously [67–69].

We will first see some examples of the first group. There are many options to assemble identified peptides to a list of proteins [24, 28, 29, 60, 64–67, 69–72]. However, statistical models are considered as a standard and preferred option [24, 29, 64–67, 69, 70]. There are many benefits to use statistical models for protein inference. First, statistical models can integrate the probabilities of peptide identification into protein inference. This can help to recover the lost connection between peptides and proteins in the digestion phase. Secondly, a natural advantage of this method is that it provides protein inference with statistical analysis. This analysis

is necessary and very important, because there is a small chance to validate the results according to any theoretical inference, since there are many variations of the internal chemical and physical process of protein digestion, peptide ionization and fragmentation. Third, there are many flexibilities in using statistical models. As is known, many factors govern the outcomes of proteome experiments. Thus, we always want to consider as many factors as possible in order to make an accurate protein inference. Meanwhile, statistical models allow us to integrate any significant factor that can decide the inference results. Furthermore, options in the model parameters can be provided so that users can apply their expertise in a specific scenario. Due to these obvious advantages, many statistical models have been proposed for protein inference. Here, one typical model is introduced to exemplify the procedure of this method, and a brief introduction of other methods is also provided.

Nesvizhskii *et al.* proposed the first statistical model for protein inference, which is implemented in the software ProteinProphet [24]. ProteinProphet infers proteins using the peptide identification probabilities produced by PeptideProphet [23]. The model is

$$P_n = 1 - \prod_{i=1}^{M_n} (1 - w_i^n p_i^n) \tag{2.7}$$

where P_n is the probability of protein n, and M_n is the number of peptides assigned to protein n. The w_i^n is the weight of peptide i being assigned to protein n, and p_i^n is the probability of peptide i being correctly identified given it is from protein n.

If peptide i has N_i parent proteins, then

$$w_i^n = \frac{P_n}{\sum\limits_{j=1,\ldots,N_i} P_j} \quad \text{and} \sum_{n=1,\ldots,N_i} w_i^n = 1. \tag{2.8}$$

It can be seen that the weight w_i^n is decided by the probability of protein n among all the parent proteins of peptide i. The probability p_i^n is computed by considering both the search engine information D_i and its number of sibling peptides (NSP) S_i^n in protein n, and is given by

$$p_i^n = p(+|D_i, \mathrm{S}_i^n) = \frac{p(+|D_i)p(\mathrm{S}_i^n|+)}{p(+|D_i)p(\mathrm{S}_i^n|+) + p(-|D_i)p(\mathrm{S}_i^n|-)}. \tag{2.9}$$

The number of sibling peptides of peptide i in protein n is written as

$$\mathrm{S}_i^n = \sum_{\{m|m\neq i\}} p(+|D_m),$$

where $p(+|D_m)$ is the probability of peptide m also from protein n given its information D_m. Information D is provided by searching engines, including matching scores and other useful information. From the formula above we can see that the estimated S_i^n is usually not an integer, because it is not really the number of sibling peptides but the sum of their identification probabilities.

In this model, degenerate peptide i is assigned to each of its parent protein $n, n = 1, \ldots, N_i$ with a weight w_i^n, which assumes that all peptides are from only one protein. These weights are computed iteratively using an expectation-maximization (EM) algorithm, as is the protein probability P_n. If one protein probability is getting higher and higher, then the weight to this protein is also becoming higher and higher. Besides, peptide identification probabilities are adjusted to integrate NSP. It shows that correctly identified peptides tend to have more siblings from one same protein, while incorrectly identified peptides tend to be the only child of its parent protein. Thus, NSP is helpful to distinguish correct peptides from incorrect ones.

ProteinProphet has been widely used to infer proteins, but there are also some problems with its model. First, it is often not true to assume that all peptides are from only one protein, especially for higher eukaryote samples in which homologous proteins exist. Second, the model tends to overestimate the probability of protein inference, because Equation(2.7) can be interpreted as the probability of a protein existing in the sample is equal to the probability that at least one identified peptide is generated by this protein. Consequently, the false-positive rate of this model is often high.

Since Tang *et al.* proposed the concept of peptide detectability [58] which can theoretically explain the assignment of degenerate peptides, this group introduced an algorithm named Lowest-Detectability First Algorithm (LDFA) to count the number of missed peptides [73]. Missed peptides are those which are not identified by searching engines, but have the detectabilities above the "lowest detectability" of an identified peptide. Then, the protein with the smallest number of missed peptides is identified first. The algorithm

will iterate until all the identified peptides are assigned to a protein. Then these proteins are reported to be identified. In essence, this algorithm also applies the parsimony principle to report a protein list, because the protein with the smallest number of missed peptides can explain more peptides than other proteins, and is identified with a priority.

Tang and his coworkers also presented another model using a Bayesian approach to infer proteins [74]. In this model, identified peptides and their parent proteins are grouped to form protein configuration graphs, shown in Figure 2.5.

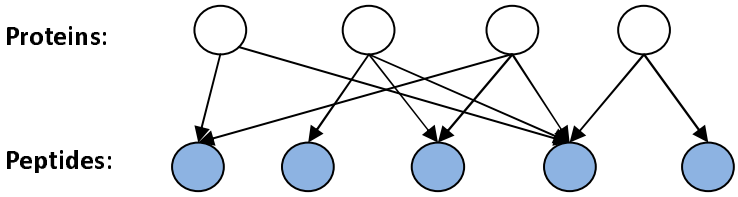


Figure 2.5: Protein configuration graph.

All the degenerate peptides and unique peptides that are assigned to a group of proteins form a peptide configuration, say $(y_1, \ldots, y_i, \ldots, y_n)$. $y_i = 1$ when the peptide i is identified; otherwise, $y_i = 0$. Then, the protein inference problem is reduced to finding the maximum a posterior (MAP) protein configuration $(x_1, \ldots, x_i, \ldots, x_m)$, which maximizes the conditional probability $P(x_1, \ldots, x_m | y_1, \ldots, y_n)$. That is,

$$(x_1, \ldots, x_m) = \underset{(x'_1, \ldots, x'_m)}{\arg\max}\, P(x_1, \ldots, x_m | y_1, \ldots, y_n), \tag{2.10}$$

where $x_i = 1$ if protein i is present, and $x_i = 0$ otherwise. The conditional probability is

$$\begin{aligned} P(x_1, \ldots, x_m | y_1, \ldots, y_n) &= \frac{P(x_1, \ldots, x_m) P(y_1, \ldots, y_n | x_1, \ldots, x_m)}{\sum_{(x'_1, \ldots, x'_m)} \left(P(x_1, \ldots, x_m) P(y_1, \ldots, y_n | x_1, \ldots, x_m) \right)} \\ &= \frac{P(x_1, \ldots, x_m) \prod_j \left[1 - P(y_j = 1 | x_1, \ldots, x_m) \right]^{1-y_j} P(y_j = 1 | x_1, \ldots, x_m)^{y_j}}{\sum_{(x'_1, \ldots, x'_m)} P(x_1, \ldots, x_m) \prod_j \left[1 - P(y_j = 1 | x_1, \ldots, x_m) \right]^{1-y_j} P(y_j = 1 | x_1, \ldots, x_m)^{y_j}}, \end{aligned} \tag{2.11}$$

in which $P(x_1, \ldots, x_m)$ is the prior probability for protein configuration. Suppose proteins are independent of each other, then

$$P(x_1, \ldots, x_m) = \prod_i P(x_i). \tag{2.12}$$

In addition,

$$P(y_j = 1|x_1, \ldots, x_m) = 1 - \prod_i [1 - x_i P\,(y_j = 1|x_i = 1, x_j = 0, j \neq i \text{ and } 1 \leq j \leq m)], \tag{2.13}$$

where $P(y_j = 1|x_i = 1, x_j = 0, j \neq i \text{ and } 1 \leq j \leq m)$ is the probability of peptide j to be identified if only protein i is present in the sample. According to the definition of peptide detectability, this is the detectability of peptide j if it comes from protein i, denoted by d_{ij}. Substituting Equations (2.12) and (2.13) into Equation (2.11) leads to

$$\begin{aligned} &P(x_1, \ldots, x_m|y_1, \ldots, y_n) \\ &= \frac{\prod_i P(x_i) \prod_j \left\{ \left[\prod_i (1 - x_i d_{ij}) \right]^{1-y_j} \left[1 - \prod_i (1 - x_i d_{ij}) \right]^{y_j} \right\}}{\sum_{(x'_1, \ldots, x'_m)} \prod_i P(x'_i) \prod_j \left\{ \left[\prod_i (1 - x'_i d_{ij}) \right]^{1-y_j} \left[1 - \prod_i (1 - x'_i d_{ij}) \right]^{y_j} \right\}}. \end{aligned} \tag{2.14}$$

This Bayesian model is solved by Gibbs Sampling. In addition to this basic model, the authors also proposed an advanced model which incorporates the peptide identification scores into the Bayesian model. Further details can refer to [74]. Because the peptide detectability is also affected by protein concentration in the sample, it needs to be converted to reflect different protein abundances. By applying this model to each protein configuration graph, all the proteins in the sample will be identified.

This model is cogent in theory because it strongly connects peptide identification with protein inference through the concept of peptide detectability and Bayes' theorem. To some extent, it addressed the problem of degenerate peptides with peptide detectability. However, there is still some room to improve this model. First, the prior probability of proteins can be refined. Secondly, more effort is needed to accurately predict peptide detectability in samples with proteins of different concentrations. Last, the model tries to identify proteins as a group. Although it mentions that marginal probability of a single protein could be computed, it doesn't solve this problem explicitly.

In addition, Higdon *et al.* proposed to use logistic function to predict proteins by a single peptide match [70]. Logistic function is very useful and flexible in predicting the presence of proteins and peptides. It can "digest" and "absorb" any factor and still produce a probability value. Specifically, if we can quantify the factors

which govern the protein digestion and peptide ionization, we can combine all of them into the logistic function and use it to identify peptides and proteins. In addition to logistic function, Shen *et al.* presented a hierarchical statistical model for protein inference [69]. Price *et al.* proposed to use Poisson distribution to simulate the distribution of the number of correct and incorrect peptide assignments [64], and then use EM algorithm to obtain the protein identification probability.

Aside from statistical models, Zhang *et al.* [60] took advantage of the mapping relationship between peptides and proteins, and adopted graph theory to infer proteins from identified peptides. They simulated the peptide-protein relationships in a bipartite graph, and employed a greedy set covering algorithm to derive a minimal protein list according to parsimony principle. An open-source software called IDPicker is developed to implement this method. A remarkable benefit of this software is that it provides the visualization of bipartite graphs, which can clearly demonstrate the peptide-protein mapping relationships, and greatly improve the transparency of protein inference. Moreover, it also reports the maximum protein list. This can help researchers with expertise to adjust the final inference results. However, the derivation of a minimal protein list leads to a conservative protein inference by nature, and it leaves out meaningful proteins from time to time. Other approaches of protein inference can refer to [28, 71, 72, 75].

Besides the methods introduced above, several other methods have been presented to identify proteins and peptides simultaneously [67–69]. Recently, Spivak *et al.* have built a Barista model [68] which formulates the protein inference as an optimization problem, shown in Figure 2.6. The protein inference problem is represented as a tripartite graph, with layers corresponding to spectra, peptides and proteins. The input to Barista is the tripartite graph with a set of features describing the match between peptides and spectra. The parameters in the model are estimated by training the model with reference data, and then the trained model is used to infer proteins. The advantage of this model is that it utilizes the spectrum information in all the steps of protein inference, without discarding spectra from peptide identification to protein inference. The application of this method is limited by the necessity of reference data to train the model each time when different datasets are analyzed.

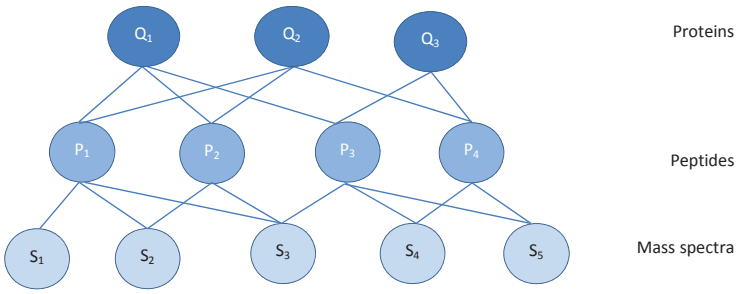


Figure 2.6: Barista tripartite graph. The tripartite graph represents the protein inference problem. From bottom to top, each layer denotes mass spectra, peptides and proteins, respectively. Barista computes a non-linear function on each PSM feature vector. Each peptide score is the maximum PSM score, and each protein score is a normalized sum of its constituent peptide scores.

Since many well-developed search engines for peptide identification are available, methods for processing peptide identification reports from these engines have been proposed. For example, Li *et al.* have used a nested mixture model [67] to estimate peptide and protein probability at the same time based on identified peptides and their scores from search engines. This model allows evidence feedback between proteins and their constituent peptides. It is built on several reasonable assumptions except that it completely ignores the problem of degenerate peptides.

2.5.3 Modifications

Post-translational modifications (PTMs) are covalent processing events that change the properties of a protein by proteolytic cleavage or by adding a modifying group to one or more amino acids [6]. PTMs of a protein can determine its activity state, localization, turnover and interactions with other proteins. Thus, identifying the modifications of a protein is an important aspect of protein characterization. Modification analysis is usually done by comparison of experimental data to known amino acid sequences [6, 8]. That is, protein identity is known and the focus is to find the modifications that this protein may carry out. Therefore, the procedure for MS-based modification analysis can be performed with the following steps [6]:

- Protein identification is conducted with MS analysis, and only unmodified peptides are considered in a database search. This can form a small database of known proteins.

- Modifications are then taken into account by searching this small known protein database. In addition, a second protease may be applied and another MS experiment can be performed in order to improve the coverage of sequences, and also increase the identification of modifications in the sample.

The confidence of identifying modified peptides is often lower than unmodified peptides, because they are searched against a much larger number of peptides. When a few modifications are considered, this problem is more significant.

2.6 Summary

Proteomics is the large-scale study of proteins, and the core instrument in proteomics is the mass spectrometer. This chapter introduced some main considerations one needs to take in designing proteomic experiments, and some often used techniques in MS-based proteomics. Methods of choice for peptide identification and protein inference were reviewed, and the challenges arising from these computational steps and possible solutions were also discussed.

Protein inference is a critical computational step in proteomics, from which the identification results serve as the foundation for further protein characterization and functional analysis. High-throughput protein inference is made convenient by MS analysis and the availability of many public genomic databases. So far, there is no perfect way to solve the protein inference problem. Although statistical models and graph theory are very good attempts, there is much space to improve these methods. First, the internal chemico-physical process of protein digestion and peptide ionization is not totally clear to us. The factors in these processes that determine the cleavage sites of proteins, the ionization ability and the charge states of peptides are not always predictable. If these factors can be quantified and included in the statistical models, the inference

accuracy should be improved. Secondly, a theoretically cogent and practically feasible concept is needed to recover the connection between peptides and proteins. Although peptide detectability is a good concept to this end, it is limited by the necessary control of protein concentration in the sample. We also proposed an MS/MS intensity-based strategy to address this problem, but this method is not yet verified with real complex proteomics data. Its practical use cannot be determined at this point. Thirdly, the identification of modified peptides in a database search is not optimized. Although database searching for modifications is possible, it is extremely paralyzed by the exponential growth of search space caused by the combinatorial explosion of modification possibilities. Last but not least, consistent validation methods are expected to analyze protein inference results, because there are no theoretical results available for reference, while the proteins or gene products vary a lot from sample to sample.

References

[1] A. Pandey and M. Mann, "Proteomics to study genes and genomics," *Nature*, 405: 837-846, 2000.

[2] T. Wehr, "Top-down versus bottom-up approaches in proteomics," *LCGC*, 2006.

[3] D. R.M. Graham, S.T. Elliott, and E. Van Eyk, "Broad-based proteomic strategies: a practical guide to proteomics and functional screening," *J. Physiol.*, 563(1): 1-9, 2005.

[4] A. Keller, S. Purvine, A. I. Nesvizhskii, S. Stolyar, D. R. Goodlett and E. Kolker, "Experimental protein mixture for validating tandem mass spectral analysis," *OMICS*, 6(2): 207-212, 2002.

[5] J. Klimek, J. S. Eddes, L. Hohmann, J. Jackson, A. Peterson, S. Letarte, P. R. Gafken, J. E. Katz, P. Mallick, H. Lee, A. Schmidt, R. Ossola, J. K. Eng, R. Aebersold and D. B. Martin, "The standard protein mi database: a diverse data set to assist in the production of improved peptide and protein identification software tools," *J. Proteome Res.*, 7: 96-103, 2008.

[6] M. Mann, and O. N. Jensen, "Proteomic analysis of post-translational modifications," *Nat. Biotechnol.*, 21: 255-261, 2003.

[7] D. A. Stead, N. W. Paton, P. Missier, S. M. Embury, C. Hedeler, B. Jin and A. J.P. Brown, "Information quality in proteomics," *Brief. Bioinform.*, 9(2): 174-188, 2008.

[8] I. Eidhammer, K. Flikka, L. Martens and S.-O. Mikalsen. Computational Methods for Mass Spectrometry Proteomics. Wiley. 2007.

[9] M. P. Washburn, "Large-scale analysis of the yeast proteome by multidimentional protein idetification technology," *Nat. Biotechnol.*, 19: 242-247, 2001.

[10] R. Aebersold and D. R. Goodlett, "Mass spectrometry in proteomics", *Chem. Rev.*, 101: 269-295, 2001.

[11] E. Kolker and R. Higdon and J. M. Hogan, "Protein idetification and expression analysis using mass spectrometry", *Trends Microbiol.*, 145: 229-235, 2006.

[12] J. Ding, J. Shi, G. G. Poirier, and F. X. Wu, "A novel approach to denoising ion trap tandem mass spectra," *BMC Proteome Science*, 7:9, 2009.

[13] W. Lin, F. X. Wu, J. Shi, J. Ding, and W. Zhang, "An adaptive approach to denoising tandem mass spectra," *Proteomics*, 11: 3773-3778, 2011.

[14] J. Shi, and F. X. Wu, "Peptide charge state determination of tandem mass spectra from low-resolution collision induced dissociation," *BMC Proteome Science*, 9(Suppl 1):S3, 2011.

[15] J. Ding, J. Shi, and F. X. Wu, "SVM-RFE based feature selection for tandem mass spectrum quality assessment," *Int. J. Data Min. Bioinform.*, 5(1): 73-88, 2011.

[16] A.M. Zou, J. Shi, J. Ding and F. X. Wu, "Charge state determination of peptide tandem mass spectra using support vector machine (SVM)," *IEEE Trans. Inf. Technol. Biomed.*, 14(3): 552-558, 2010.

[17] B. M. Webb-Robertson and W. R. Cannon, "Current trends in computational inference from mass spectrometry-based proteomics," *Bioinformatics*, 8: 304-317, 2007.

[18] D. N. Perkins, D. J. C. Pappin, D. M. Creasy and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, 20: 3551-3567, 1999.

[19] J. K. Eng, A. L. McCormack and J. R. Yates III, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *J. Am. Soc. Mass Spectrom.*, 5: 976-989, 1994.

[20] D. Li, Y. Fu, R. Sun, C. X. Ling, Y. Wei, H. Zhou, R. Zeng, Q. Yang, S. He and W. Gao, "pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry," *Bioinformatics*, 21: 3049-3050, 2005.

[21] R. Craig and R. C. Beavis. "TANDEM: matching proteins with tandem mass spectra," *Bioinformatics*, 20: 1466-1467, 2004.

[22] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Marynard, X. Yang, W. Shi, S. H. Bryant, "Open mass spectrometry search algorithm," *J. Proteome Res.*, 3: 958-964, 2004.

[23] A. Keller, A. I. Nesvizhskii, E. Kolker and R. Aebersold,"Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search," *Anal. Chem.*, 74: 5383-5392, 2002.

[24] A. I. Nesvizhskii, A. Keller, E. Kolker and R. Aebersold, "A statistical model for identifying proteins by tandem mass spectrometry," *Anal. Chem.*, 75: 4646-4658, 2003.

[25] J. Shi, W. Lin, and F.-X. Wu, "Statistical analysis of Mascot peptide identification with active logistic regression," *iCBBE*, 2010.

[26] B. Kuster, M. Schirle, P. Mallick and R. Aebersold, "Scoring proteomes with proteotypic peptide probes," *Nat. Rev. Mol. Cell Biol.*, 6: 577-583, 2005.

[27] H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, S. E. Stein and R. Aebersold, "Development and validation of a spectral library searching method for peptide identification from MS/MS," *Proteomics*, 7: 655-667, 2007.

[28] R. Craig, J. P. Cortens and R. C. Beavis, "The use of proteotypic peptide libraries for protein identification," *Rapid Commun. Mass Spectrom.*, 19: 1844-1850, 2005.

[29] J. Feng, D. Q. Naiman and B. Cooper, "Probability model for assessing proteins assembled from peptide sequences inferred from tandem mass spectrometry data," *Anal. Chem.*, 79: 3901-3911, 2007.

[30] Y. Lin, Y. Qiao, S. Sun, C. Yu, G. Dong and D. Bu. "A fragmentation event model for peptide identification by mass spectrometry," *RECOMB*, LNBI 4955: 154-166, 2008.

[31] J. Colinge and K. L. Bennett, "Introduction to Computational Proteomics," *PLoS Computational Biology*, 307: 1151-1160, 2007.

[32] B. M. Webb-Robertson, W. R. Cannon, C. S. Oehmen, A. R. Shah, V. Gurumoorthi, M. S. Lipton and K. M. Waters, "A support vector machine model for the prediciton of peoteotypic peptides for accurate mass and time proteotimics," *Bioinformatics*, 24: 1503-1509, 2008.

[33] A. I. Nesvizhskii and R. Aebersold, "Interpretation of shotgun proteomic data: the protein inference problem," *Mol. Cell. Proteomics*, 4(10): 1419-1440, 2005.

[34] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, 422: 198-207, 2003

[35] M. A. Baldwin, "Protein identification by mass spectrometry," *Mol. Cell. Proteomics*, 3: 1-9, 2004.

[36] J. A. Taylor and R. S. Johnson, "Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry," *Anal. Chem.*, 73: 2594-2604, 2001.

[37] J. A. Taylor and R. S. Johnson, "Sequence database searches via de novo peptide sequencing by tandem mass spectrometry," *Rapid Commun. Mass Spectrom.*, 11: 1067-1075, 1997.

[38] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby and G. Lajoie, "PEAKS: Powerful software for peptide de novo sequecning by tandem mass spectrometry," *Rapid Commun. Mass Spectrom.*, 17: 2337-2342, 2003.

[39] B. Ma, K. Zhang, and C. Liang, "An effective algorithm for peptide de novo sequencing from MS/MS spectra," *JCSS*, 70: 418-430, 2005.

[40] A. Frank and P. Pevzner, "PepNovo: De novo peptide sequencing via probabilistic network modeling," *Anal. Chem.*, 77: 964-973, 2005.

[41] T. Chen, M. Y. Kao, M. Tepel, J. Rush and G. M. Church, "A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry," *J. Comp. Biol.*, 8: 325-337, 2001.

[42] L. Mo, D. Dutta, Y. Wan and T. Chen, "MSNovo: A dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry," *Anal. Chem.*, 79: 4870-4878, 2007.

[43] C. Xu and B. Ma, "Software for computational peptide identification from MS/MS data," *Drug Discov. Today*, 11: 595-600, 2006.

[44] B. Lu and T. Chen, "Algorithms for de novo peptide sequencing using tandem mass spectrometry," *Drug Discov Today Biosilico*, 2: 85-90, 2004.

[45] R. E. Moore, M. K. Young and T. D. Lee, "Qscore: an algorithm for evaluating SEQUEST database search results," *J. Am. Soc. Mass Spectrom.*, 13: 378-386, 2002.

[46] P. A. Rudnick, Y. Wang, E. Evans, C. S. Lee and B. M. Balgley, "Large scale analysis of MASCOT results using a mass accuracy-based threshold (MATH) effectively improves data interpretation," *J. Proteome Res.*, 4: 1353-1360, 2005.

[47] J. E. Elias and S. P Gygi, "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry," *Nat. Methods*, 4: 207-214, 2007.

[48] L. Kall, J. D. Storey, M. J. MacCoss, and W. S. Noble, "Assigning significance to peptides identified by tandem mass spectrometry using decoy databases," *J. Proteome Res.*, 7: 29-34, 2008.

[49] H. Choi, D. Ghosh, and A. I. Nesvizhskii, "Assigning significance to peptides identified by tandem mass spectrometry using decoy databases," *J. Proteome Res.*, 7: 47-50, 2008.

[50] H. Choi, and A. I. Nesvizhskii, "False discovery rates and related statistical concepts in mass spectrometry-based proteomics," *J. Proteome Res.*, 7: 29-34, 2008.

[51] B. Thiede, W. Hohenwarter, A. Krah, J. Mattow, M. Schmid, F. Schmidt and P. R. Jungblut, "Peptide mass fingerprinting," *Methods*, 35: 237-247, 2005.

[52] A. E. Ashcroft, "Protein and peptide identification: the role of mass spectrometry in proteomics," *Nat. Prod. Rep.*, 20: 202-215, 2003.

[53] B. Domon and R. Aebersold, "Mass spectrometry and protein analysis," *Science*, 312: 212-217, 2006.

[54] J. Shi, B. Chen and F. X. Wu, "Improve accuracy of peptide identification with consistency between peptides," *IEEE BIBM*, 191-196, 2011.

[55] Z. He, H. Zhao and W. Yu, "Score regularization for peptide identification," *BMC Bioinfomatics*, 12(Suppl):S2, 2011.

[56] J. Shi, B. Chen and F. X. Wu, "Unifying protein inference and peptide identification with feedback to update consistency between peptides," *Proteomics*, 2012, accepted.

[57] J. Shi, and F. X. Wu, "A feedback framework for protein inference with peptides identified from tandem mass spectra," *Proteome Science*, 2012, accepted.

[58] H. Tang, R. J. Arnold, P. Alves, Z. Xun, D. E. Clemmer, M. V. Novotny, J. P. Reilly and P. Radivojac, "A computational approach toward label-free protein quantification using predicted peptide detectability," *Bioinformatics*, 22: e481-e488, 2006.

[59] I. J. Good, "Explicativity: a mathematical theory of explanation with statistical applications," *Proc. R. Soc.*, 354: 303-330, 1997. London.

[60] B. Zhang, M. C. Chambers and D. L. Tabb, "Proteomic parsimony through bipartite graph analysis improves accuracy and transparancy," *J. Proteome Res.*, 6: 3549-3557, 2007.

[61] W. S. Sanders, S. M. Bridges, F. M. McCarthy, B. Nanduri and S. C. Burgess, "Prediction of peptides observable by mass spectrometry applied at the experimental set level," *Bioinformatics*, 8(Suppl: 6): S12, 2007.

[62] T. Huang, J. Wang, W. Yu, and Z. He, "Protein inference: a review," *Brief. Bioinform.*, 2012, 13:586-614.

[63] Shi, J., Wu, F.-X., "Protein inference by assembling peptides identified from tandem mass spectra," *Curr. Bioinform.*, 4: 226-233, 2009.

[64] T. S. Price, M. B. Lucitt, W. Wu, D. J. Austin, A. Pizarro, A. K. Yocum, I. A. Blair, G. A. FitzGerald and T. Grosser, "EBP: a program for protein identification using multiple tandem mass spectrometry datasets," *Mol. Cell. Proteomics*, 6: 527-536, 2007.

[65] Y. F. Li, R. J. Arnold, Y. Li, P. Radivojac, Q. Sheng and H. Tang, "A Bayesian approach to protein inference problem in shotgun proteomics," *J. Comput. Biol.*, 16:1183-1193, 2009.

[66] P. Alves, R. J. Arnold, M. V. Novotny, P. Radivojac, J. P. Reilly and H. Tang, "Advancement in protein inference from shotgun proteomics using peptide detectability," *Pac. Symp. Biocomput.*, 12: 409-420, 2007.

[67] Q. Li, M. MacCoss and M. Stephens, "A nested mixture model for protein identification using mass spectrometry," *Ann. Appl. Stat.*, 4(2): 962-987, 2010.

[68] M. Spivak, D. Tomazela, J. Weston, M. J. MacCoss, and W. S. Noble, "Direct maximization of protein identifications from tandem mass spectra," *Mol. Cell. Proteomics*, 2012.

[69] C. Shen, Z. Wang, G. Shankar, X. Zhang and L. Li, "A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry," *Bioinformatics*, 24: 202-208, 2007.

[70] R. Higdon and E. Kolker, "A predictive model for identifying proteins by a single peptide match," *Bioinformatics*, 23: 277-280, 2007.

[71] S. Gerster, E. Qeli, C. H. Ahrens and P. Buhlmann, "Protein and gene model inference based on statistical modeling in k-partite graphs," *PNAS*, 107(27): 12101-12106, 2010.

[72] N. Bandeira, D. Tsur, A. Frank and P. A. Pevzner, "Protein identification by spectral networks analysis," *PNAS*, 104: 6140-6145, 2007.

[73] P. Alves, R. J. Arnold, M. V. Novotny, P. Radivojac, J. P. Reilly and H. Tang, "Advancement in

protein inference from shotgun proteomics using peptide detectability," *Pac. Symp. Biocomput.*, 12: 409-420, 2007.

[74] Y. F. Li, R. J. Arnold, Y. Li, P. Radivojac, Q. Sheng and H. Tang, "A Bayesian approach to protein inference problem in shotgun proteomics," *J. Comput. Biol.*, 16:1183-1193, 2009.

[75] P. Kearney, H. Butler, K. Eng and P. Hugo, "Protein identification and peptide expression resolver: harmonizing protein identification with protein expression data," *J. Proteome Res.*, 7: 234-244, 2008.

Chapter 3

Peptide charge state determination of tandem mass spectra from low-resolution collision induced dissociation

Published as: Jinhong Shi, and Fang-Xiang Wu, Peptide charge state determination of tandem mass spectra from low-resolution collision induced dissociation, *Proteome Science*, vol.9(Suppl 1):S3, 2011. This work was first presented in *International Workshop on Computational Proteomics*, Hong Kong, China. 18-21 December 2010.

In the previous chapter, we introduced the basic concepts and principles of tandem mass spectrometry for protein inference. From MS/MS data to inferred proteins, there are three computational phases: First, process MS/MS data to improve the quality of peptide identification; Second, postprocessing peptide identification results from search engines; and third, infer proteins based on the identified peptides and their probabilities. Note that a peptide-spectrum-match is performed with existing search engines. In this thesis, we will introduce the work on each phase, respectively. This chapter will present a method to determine the charge states of peptide tandem mass spectra from low-resolution collision induced dissociation (CID), which is one aspect among various processing of MS/MS data.

The manuscript included in this chapter studies the determination of low-resolution CID tandem mass spectra with an unsupervised machine learning method, Gaussion mixture model (GMM). Four novel and

discriminant features are proposed to represent each tandem mass spectrum, and are used in GMM to distinguish doubly and triply charged peptides. The results have shown that this method is easier and more accurate to assign charge states to low-resolution tandem mass spectra than existing methods.

Peptide charge state determination of tandem mass spectra from low-resolution collision induced dissociation

Abstract

Charge states of tandem mass spectra from low-resolution collision induced dissociation cannot be determined by mass spectrometry. As a result, such spectra with multiple charges are usually searched multiple times by assuming each possible charge state. Not only does this strategy increase the overall database search time, but also yields more false positives. Hence, it is advantageous to determine charge states of such spectra before a database search. We propose a new approach capable of determining the charge states of low-resolution tandem mass spectra. Four novel and discriminant features are introduced to describe tandem mass spectra and used in Gaussian mixture model to distinguish doubly and triply charged peptides. By testing on three independent datasets with known validity, the results show that this method can assign charge states to low-resolution tandem mass spectra more accurately than existing methods. The proposed method can be used to improve the speed and reliability of peptide identification.

3.1 Background

Mass spectrometry has been widely used to analyze high throughput protein samples. Proteins are first cleaved into peptides with enzymes or chemical cleavages. Then, peptides are separated from mixture solutions by high pressure liquid chromatography (HPLC), and sent to an ionization source where they get ionized. There are two ionization techniques, electrospray ionization (ESI) and matrix assisted laser desorption/ionization (MALDI), which are often used in proteomics laboratories. MALDI is mainly used in peptide mass fingerprinting as it predominantly yields singly charged ions. Unlike MALDI, ESI typically produces multiply charged ions. After being ionized, peptides are introduced into analyzers such as ion trap or triple quadrupole to produce mass spectra (MS). To obtain tandem mass spectra (MS/MS), peptide ions

with the highest intensities in MS are isolated and subjected to fragmentation by collision induced dissociation (CID). The resultant MS/MS are used to provide structural composition information of peptides.

The commonly used database search programs for peptide identification include Sequest [2] and Mascot [3]. These programs compare experimental spectra with theoretical spectra in a database and use scoring functions to measure the similarity between them. Typically, the peptide with the highest score is identified. However, the growing number of protein sequences in expanding databases becomes a challenge for database search software because the search space is sharply increasing. Moreover, multiply charged peptide tandem mass spectra from ESI-CID also add complexities to these programs, because they generate much more complex spectra. Although high-resolution mass spectrometers can provide separable isotropic spacing of fragment ions to derive charge states, most commonly used ion trap and triple quadrupole analyzers have limited resolution to do so [4]. In such a case, one spectrum is usually searched multiple times by assuming each possible charge state of its precursor peptide ion. This strategy increases the overall time of database search and yields more false positives as true positives need to be distinguished from much more peptide candidates. The requirement of determining peptide charge states is not limited to database searching, but also is necessary in de novo sequencing methods [5].

This paper will focus on the charge state determination of low-resolution tandem mass spectra. There have been reports in determining charge states of low-resolution tandem mass spectra [1, 4, 6, 7]. Thirty-four features were proposed in [6] to describe MS/MS and the link between MS and MS/MS, then a support vector machine (SVM) was used to classify MS/MS into three groups +2, +3 and +2/+3. One problem with this method is that it classifies peptide ions into three groups, which still leaves ambiguities in the charge determination. Lately, twenty-eight features of MS/MS were proposed to train an SVM in [7] to discriminate doubly and triply charged peptides. The common problem with [6, 7] is that an SVM needs training with labeled data (labeled data are the ones that we know to which class they belong). This inherent drawback of supervised methods limits their generality in determining the charges of any experimental MS/MS. Last but not least, it is computationally expensive to first train SVM and then apply it on test data.

In this paper, we present an unsupervised learning method based on Gaussian mixture model (GMM) to determine the charge states of low-resolution tandem mass spectra. Four novel and discriminant features are proposed to describe MS/MS. By testing on three low-resolution MS/MS datasets with verified charge states, the results have shown that the proposed method can accurately assign charge states to such tandem mass spectra.

3.2 Methods

In a database search, tandem mass spectra are usually considered to carry 1, or 2 or 3 charges. Research [8] shows that singly charged MS/MS can be reliably determined. Therefore, the charge state determination can be reduced to the classification of doubly and triply charged MS/MS. To solve this problem, this study uses the unsupervised GMM with features proposed to reflect the properties of MS/MS. Since the features are to be extracted from MS/MS, we will first introduce several properties of peptide CID tandem mass spectra. For more details about these properties, we would refer readers to [9].

Properties of CID tandem mass spectra

Let $m(\mathrm{a}_i)$ be the mass of amino acid a_i, then the mass of peptide P with n amino acids is given by

$$m(\mathrm{P}) = m(\mathrm{H}) + \sum_{i=1}^{n} m(\mathrm{a}_i) + m(\mathrm{OH}) \tag{3.1}$$

where $m(\mathrm{H})$ and $m(\mathrm{OH})$ are the masses of the additional N-terminal and C-terminal. The cleavage along peptide bonds in CID mainly leads to the production of N-terminal b_i ion and C-terminal y_{n-i} ion. The singly charged ion with N-terminal is denoted by b_i^+, and its m/z value is

$$m(\mathrm{b}_i^+) = m(\mathrm{H}) + \sum_{j=1}^{i} m(\mathrm{a}_j). \tag{3.2}$$

The m/z value of its doubly charged counterpart b_i^{++} is

$$m(\mathrm{b}_i^{++}) = [m(\mathrm{b}_i^{+}) + m(\mathrm{H})]/2. \tag{3.3}$$

The singly charged ion with C-terminal is denoted by y_{n-i}^{+}, and its m/z value is

$$m(\mathrm{y}_{n-i}^{+}) = 2 * m(\mathrm{H}) + m(\mathrm{OH}) + \sum_{j=i+1}^{n} m(\mathrm{a}_j). \tag{3.4}$$

Here two hydrogens are added because C-terminal ions carry one negative charge after fragmentation, thus it needs two protons to make it carry one positive charge. Similarly, the m/z value of its doubly charged counterpart y_{n-i}^{++} is

$$m(\mathrm{y}_{n-i}^{++}) = [m(\mathrm{y}_{n-i}^{+}) + m(\mathrm{H})]/2. \tag{3.5}$$

From equations (3.1) to (3.5), we have the following equations holding for peptide CID tandem mass spectra:

$$m(\mathrm{P}) + 2 * m(\mathrm{H}) = m(\mathrm{b}_i^{+}) + m(\mathrm{y}_{n-i}^{+}) \tag{3.6}$$

$$m(\mathrm{P})/2 + 2 * m(\mathrm{H}) = m(\mathrm{b}_i^{++}) + m(\mathrm{y}_{n-i}^{++}) \tag{3.7}$$

$$m(\mathrm{P})/2 + 2 * m(\mathrm{H}) = m(\mathrm{b}_i^{++}) + (m(\mathrm{y}_{n-i}^{+}) + m(\mathrm{H}))/2 \tag{3.8}$$

$$m(\mathrm{P})/2 + 2 * m(\mathrm{H}) = (m(\mathrm{b}_i^{+}) + m(\mathrm{H}))/2 + m(\mathrm{y}_{n-i}^{++}). \tag{3.9}$$

Since one peptide with different charges can produce different MS/MS, we can infer the charge state of a peptide according to the features of its MS/MS. As we will see, these features will be calculated based on the above relationships between the singly and doubly charged fragment ions.

Spectrum features

First, six functions are defined for a given peptide MS/MS [9] as follows:

$$d_1(m_1, m_2) = m_2 - m_1$$

$$s_1(m_1, m_2) = m_1 + m_2$$

$$d_2(m_1, m_2) = m_2 - (m_1 + 1)/2$$

$$d_3(m_1, m_2) = (m_2 + 1)/2 - m_1$$

$$s_2(m_1, m_2) = m_1 + (m_2 + 1)/2$$

$$s_3(m_1, m_2) = (m_1 + 1)/2 + m_2$$

where m_1 and m_2 are the m/z values of any two peaks from the given peptide tandem mass spectrum and $m_2 > m_1$.

Complementary pairs

Complementary pairs measure the likelihood that an N-terminal ion and a C-terminal ion in a peptide MS/MS are produced as the peptide fragments at the same peptide bond. Given a peptide P and MS/MS data, let

$$\mathcal{S}_1 = \{(m_1, m_2) \mid s_1(m_1, m_2) \approx m(\mathrm{P}) + 2 * m(\mathrm{H}), m_1, m_2 \in .., m_1 < m_2.\}$$

$$\mathcal{S}_2 = \{(m_1, m_2) \mid s_2(m_1, m_2) \approx m(\mathrm{P})/2 + 2 * m(\mathrm{H}), m_1, m_2 \in .., m_1 < m_2.\}$$

$$\mathcal{S}_3 = \{(m_1, m_2) \mid s_3(m_1, m_2) \approx m(\mathrm{P})/2 + 2 * m(\mathrm{H}), m_1, m_2 \in .., m_1 < m_2.\}$$

then, the first feature is defined as

$$\delta_{\mathrm{cp}} = |\mathcal{S}_1| - (|\mathcal{S}_2| + |\mathcal{S}_3|) \tag{3.10}$$

where $|\cdot|$ denotes the cardinality of a set. The feature δ_{cp} is the difference between the number of complementary pairs (+1, +1) and the number of complementary pairs (+1, +2) in MS/MS. This feature accounts for the fact that +2 peptides tend to generate two +1 ions at the same bond, while +3 peptides are prone to yield one +1 and one +2 ion [1, 4]. From the definition, this feature is expected to be larger for doubly charged peptides than triply charged ones.

According to the definition of s_1, s_2 and s_3, we define peak sets

$$\mathcal{P}_{11}^{+} = \{m_1 \mid (m_1, m_2) \in \mathcal{S}_1\}, \quad \mathcal{P}_{12}^{+} = \{m_2 \mid (m_1, m_2) \in \mathcal{S}_1\}$$

$$\mathcal{P}_{2}^{++} = \{m_1 \mid (m_1, m_2) \in \mathcal{S}_2\} \ \cup \ \{m_2 \mid (m_1, m_2) \in \mathcal{S}_3\}$$

$$\mathcal{P}_{2}^{+} = \{m_1 \mid (m_1, m_2) \in \mathcal{S}_3\} \ \cup \ \{m_2 \mid (m_1, m_2) \in \mathcal{S}_2\}.$$

Then, the second feature is given by

$$\delta_{\mathrm{R_{cp}}} = \frac{\sum\limits_{m \in \mathcal{P}_{12}^{+}} I(m)}{0.5 + \sum\limits_{m \in \mathcal{P}_{11}^{+}} I(m)} - \frac{\sum\limits_{m \in \mathcal{P}_{2}^{++}} I(m)}{0.5 + \sum\limits_{m \in \mathcal{P}_{2}^{+}} I(m)} \tag{3.11}$$

where $I(\cdot)$ represents the intensity of peaks. The feature $\delta_{\mathrm{R_{cp}}}$ is the difference between the ratio of $+1$ peak intensity over their complementary $+1$ peak intensity and the ratio of $+2$ peak intensity over their complementary $+1$ peak intensity. The item 0.5 is added in view that the intensity of y ions in higher mass regions is larger than that of b ions in lower mass regions. This feature accounts for the fact that the intensity of $+1$ peaks and the intensity of their complementary $+1$ peaks should be comparable when they are produced from doubly charged peptides, while the intensity of $+1$ peaks from triply charged peptides should be comparable to the intensity of their complementary $+2$ peaks. Thus, the difference between these two ratios should be greater than 0 for doubly charged peptides while less than 0 for triply charged ones. This newly proposed feature is expected to be more significant than the first feature which has been used in [4], because it integrates the intensity information into the feature definition rather than just counting the number of complementary pairs.

Regional intensity

Intensity is an important property of tandem mass spectra, so we incorporate it into the expression of the third feature. Let

$$\mathcal{D}_1 = \{(m_1, m_2) \mid d_1(m_1, m_2) \approx M_i/2, i = 1, 2 \ldots 20\}$$

$$\mathcal{D}_2 = \{(m_1, m_2) \mid d_2(m_1, m_2) \approx M_i/2, i = 1, 2 \ldots 20\}$$

$$\mathcal{D}_3 = \{(m_1, m_2) \mid d_3(m_1, m_2) \approx M_i/2, i = 1, 2 \ldots 20\},$$

where M_i is the residue mass of the amino acid i. Then according to the definition of d_1, d_2, d_3, we can see that the set of doubly charged peaks is

$$\mathcal{P}^{++} = \{m_1 \mid (m_1, m_2) \in \mathcal{D}_1\} \cup \{m_2 \mid (m_1, m_2) \in \mathcal{D}_1\} \cup \{m_2 \mid (m_1, m_2) \in \mathcal{D}_2\} \cup \{m_1 \mid (m_1, m_2) \in \mathcal{D}_3\}.$$

In view of further manipulation, we define an indicator function of the peak masses in a spectrum,

$$X(m) = \begin{cases} 1 & m \in [m_p, 1.5m_p] \\ 0 & \text{otherwise} \end{cases}$$

where m_p is the m/z value of parent peptide ions. Then the third feature is defined as

$$I_{\text{dc}} = \sum_{m \in \mathcal{P}^{++}} I(m)X(m). \tag{3.12}$$

The feature I_{dc} is the intensity of +2 peaks in the mass region $[m_p, 1.5m_p]$. In theory, the m/z values of +2 peaks from +2 peptides should not exceed m_p, while they should not exceed $1.5m_p$ when they are from +3 peptides. Hence, I_{dc} which accounts for the +2 peak intensity in the region $[m_p, 1.5m_p]$ should be very discriminant for doubly and triply charged peptides. This feature is expected to be smaller for doubly charged peptides than triply charged ones.

Amino acid distance

The charge state of a peptide is theoretically determined by the number of basic amino acids it contains [10]. The side chains of basic sites have high proton affinities to attract protons in ESI, and the N-terminal amine group can also attract a proton. Thus in theory, doubly charged peptides should contain one basic site and triply charged peptides should contain two basic sites. Let n_{bs} be the number of basic sites of an MS/MS, and define

$$\mathcal{D} = \{(m_1, m_2) | d_1(m_1, m_2) \approx M_a,\ a = \text{K}, \text{R}, \text{H}\} \cup \{(m_1, m_2) | d_1(m_1, m_2) \approx M_a/2,\ a = \text{K}, \text{R}, \text{H}\}$$

$$\{(m_1, m_2) | d_2(m_1, m_2) \approx M_a/2,\ a = \text{K}, \text{R}, \text{H}\} \cup \{(m_1, m_2) | d_3(m_1, m_2) \approx M_a/2,\ a = \text{K}, \text{R}, \text{H}\},$$

where M_a is the residue mass of the amino acid a. Then the number of basic sites is computed by

$$n_{\text{bs}} = |\mathcal{D}| \ / \ N_t, \tag{3.13}$$

where N_t is the theoretical repeat number of basic residues in a mass spectrum. More discussion about n_{bs} is given later.

When we compute the values of all features, the situations when peaks are produced by losing water, ammonia, CO or NH group are considered as proposed in [7].

3.2.1 Gaussian mixture model

Gaussian mixture model (GMM) [11] is commonly used for clustering and it is unsupervised, which makes GMM have an obvious advantage over other supervised methods in terms of saving efforts in labeling training data. The expression of Gaussian mixtures is given by

$$f(\mathbf{x};\theta) = \sum_{k=1}^{K} p_k g(\mathbf{x};\mu_k,\sigma_k) \tag{3.14}$$

where

$$g(\mathbf{x};\mu_k,\sigma_k) = \frac{1}{(\sqrt{2\pi}\sigma_k)^D} e^{-\frac{1}{2}\left(\frac{\|\mathbf{x}-\mu_k\|}{\sigma_k}\right)^2}, \tag{3.15}$$

$\|\|$ is 2-norm of a vector, and p_k is the mixing probability of the k^{th} component. Here, D is the space dimension of data points. The maximum likelihood approach is used to estimate the parameter vector θ in GMM. The likelihood function is given by

$$\lambda(\mathbf{X};\theta) = \prod_{n=1}^{N} f(\mathbf{x}_n;\theta) \tag{3.16}$$

Substituting the Gaussian mixtures (3.14) into (3.16), and taking the logarithm of the likelihood function, we have

$$L(\mathbf{X};\theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} p_k g(\mathbf{x}_n;\mu_k,\sigma_k). \tag{3.17}$$

Then, the parameter θ is given by

$$\hat{\theta} = \arg\max_{\theta} L(\mathbf{X}; \theta). \tag{3.18}$$

To solve (3.18), we take the derivatives of L with respect to μ_k and σ_k, which yields

$$\frac{\partial L}{\partial \mu_k} = \sum_{n=1}^{N} \frac{p(k|n)}{\sigma_k^2}(\mu_k - \mathbf{x}_n) \tag{3.19}$$

$$\frac{\partial L}{\partial \sigma_k} = \sum_{n=1}^{N} p(k|n)(-\frac{D}{\sigma_k} + \frac{\|\mathbf{x}_n - \mu_k\|^2}{\sigma_k^3}) \tag{3.20}$$

where

$$p(k|n) = \frac{p(k,n)}{p(n)} = \frac{p(k,n)}{\sum_{z=1}^{K} p(z,n)}. \tag{3.21}$$

In the above expression, $p(k,n)$ is defined as

$$p(k,n) = p_k p(n|k) = p_k g(\mathbf{x}_n; \mu_k, \sigma_k). \tag{3.22}$$

To obtain the derivative of L with respect to the mixing probability p_k, we write the variables p_k as functions of unconstrained variables γ_k [12], given in (3.23), because the values of p_k are constrained to being positive and adding up one.

$$p_k = \frac{e^{\gamma_k}}{\sum_{z=1}^{K} e^{\gamma_z}} \tag{3.23}$$

This transform enforces both constraints automatically. From the chain rule of differentiation, we obtain

$$\frac{\partial L}{\partial \gamma_k} = \sum_{n=1}^{N} (p(k|n) - p_k). \tag{3.24}$$

Setting all derivatives to zero, we obtain three groups of equations for the means, variances, and mixing probabilities:

$$\mu_k = \frac{\sum_{n=1}^{N} p(k|n)\mathbf{x}_n}{\sum_{n=1}^{N} p(k|n)} \tag{3.25}$$

$$\sigma_k^2 = \frac{1}{D}\frac{\sum_{n=1}^{N} p(k|n)\,\|\mathbf{x}_n - \mu_k\|^2}{\sum_{n=1}^{N} p(k|n)} \tag{3.26}$$

$$p_k = \frac{1}{N}\sum_{n=1}^{N} p(k|n). \tag{3.27}$$

These equations are intimately coupled with one another, because the term $p(k|n)$ in turn depends on all terms on the left-hand sides through (3.21) and (3.22). Thus, it is hard to solve these equations directly. However, the EM algorithm can provide a solution. We start with a guess for the parameters p_k, μ_k, σ_k, and then iteratively cycle through (3.21), (3.22) (E-step), and then (3.25), (3.26) and (3.27) (M-step). The procedures of EM algorithm are given as follows:

• E-step:

$$p^{(i)}(k|n) = \frac{p_k^{(i)} g(\mathbf{x}_n; u_k^{(i)}, \sigma_k^{(i)})}{\sum_{z=1}^{K} p_z^{(i)} g(\mathbf{x}_n; \mu_z^{(i)}, \sigma_z^{(i)})} \tag{3.28}$$

• M-step:

$$\mu_k^{(i+1)} = \frac{\sum_{n=1}^{N} p^{(i)}(k|n)\mathbf{x}_n}{\sum_{n=1}^{N} p^{(i)}(k|n)} \tag{3.29}$$

$$\sigma_k^{2^{(i+1)}} = \frac{1}{D} \frac{\sum_{n=1}^{N} p^{(i)}(k|n) \left\| \mathbf{x}_n - \mu_k^{(i+1)} \right\|^2}{\sum_{n=1}^{N} p^{(i)}(k|n)} \tag{3.30}$$

$$p_k^{(i+1)} = \frac{1}{N} \sum_{n=1}^{N} p^{(i)}(k|n). \tag{3.31}$$

3.3 Results and Discussion

3.3.1 Experimental data

Three datasets are used to investigate the performance of the proposed method in predicting charge states of peptide CID tandem mass spectra.

- **ISB dataset** was acquired on an LC-ESI ion trap (ThermoFinnigan) and was provided by the Institute of Systems Biology (ISB, Seattle, USA). It contains $37,044$ peptide MS/MS spectra from a control mixture of 18 standard proteins [13]. The charge states were assigned to 1656 doubly charged and 984 triply charged peptides with Sequest.

- **TOV dataset** includes $22,577$ peptide MS/MS spectra which were acquired on an LCQ DECA XP ion trap (Thermo Electron Corp.). The samples analyzed were generated by the tryptic digestion of a whole-cell lysate from 36 fractions of TOV-112D [14]. These spectra were searched using Sequest and the assignments of 1898 doubly charged and 261 triply charged spectra were verified to be correct by Scaffold (http://www.proteomesoftware.com) with the minimum probability of 0.95.

- **BALF dataset** was obtained from an LCQ DECA ion trap mass spectrometer (ThermoFinnigan) and is available in PeptideAtlas (http://www.peptideatlas.org/repository) data repository. MS/MS were searched with Sequest against IPI human protein database. The assignments of 2492 doubly charged and 3686 triply charged spectra were validated using PeptideProphet with the minimum probability 0.90.

3.3.2 Results

GMM is solved by implementing the EM algorithm described previously with MATLAB. All features are transformed to have variances of 1. A receiver operating characteristic (ROC) curve and Area Under the Curve (AUC) are employed to measure the classifier performance. ROC curves of actual classifications locate in between the ideal plot (the point $(0, 1)$) and the random-guess plot (the diagonal line) with AUC $\in (0.5, 1)$. The bigger the AUC, the more powerful the classification is.

Comprehensive performance of the features

First, we build the classifier with all features to see their comprehensive performance. The estimated means of the four features for doubly and triply charged peptides of the three datasets are shown in Table 3.1. It can be seen that all these estimated values are consistent to the expected values. ROC curves of the three datasets are given in Figure 3.1. AUC for ISB, TOV and BALF are $0.9732, 0.9903, 0.9990$, respectively.

Both ROC and AUC show that GMM with the proposed features is well-suited for the classification of low-resolution peptide CID tandem mass spectra.

Table 3.1: Estimates of means of all features for +2 and +3 MS/MS and their expected relationships.

Features	ISB		TOV		BALF		EXPECTED
	+2	+3	+2	+3	+2	+3	Feature values
δ_{cp}	−0.0956	−1.5366	−0.4592	−2.1642	−0.8590	−2.3805	+2 > +3
$\delta_{\mathrm{R_{cp}}}$	0.8384	−0.5340	0.8842	−0.4470	0.4762	−1.3666	+2 > +3
I_{dc}	0.2099	1.4521	0.3941	2.0239	0.4743	1.5057	+2 < +3
n_{bs}	0.4887	1.4556	0.9962	2.1185	1.2003	1.2302	+2 < +3

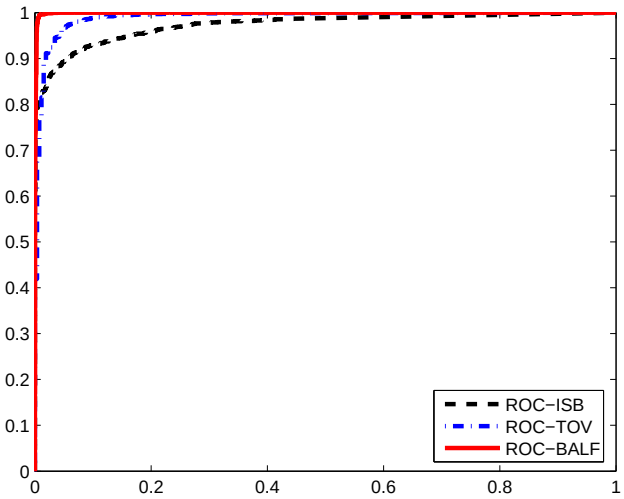


Figure 3.1: ROC curves of ISB, TOV, and BALF data with all features. $\mathrm{AUC_{ISB}} = 0.9732$, $\mathrm{AUC_{TOV}} = 0.9903$, $\mathrm{AUC_{BALF}} = 0.9990$.

Discriminant power of each feature

Here we examine the power of each proposed feature in discriminating doubly charged and triply charged peptides with AUC, which is given in Table 3.2. The AUC shows that the most significant feature is $\delta_{\mathrm{R_{cp}}}$, which measures the comparable degree of the intensity of complementary pairs. The second one is the commonly used feature δ_{cp} and the third one is I_{dc}, which accounts for the intensity difference of doubly charged peaks in the mass region $[m_p, 1.5m_p]$. The feature with the least discriminant power is the number of basic sites n_{bs}. Theoretically, this feature reflects the origin of the charges carried by peptides through ESI, thus it should be significant in distinguishing doubly and triply charged peptides. More discussions are given for this inconsistent result in the following subsection.

The three most significant features are used to build the GMM classifier and the performance is given in Figure 3.2. It is obvious that the classifier is very powerful in separating doubly charged and triply charged peptides in all three datasets. Furthermore, it is even better than the classifier built with all features.

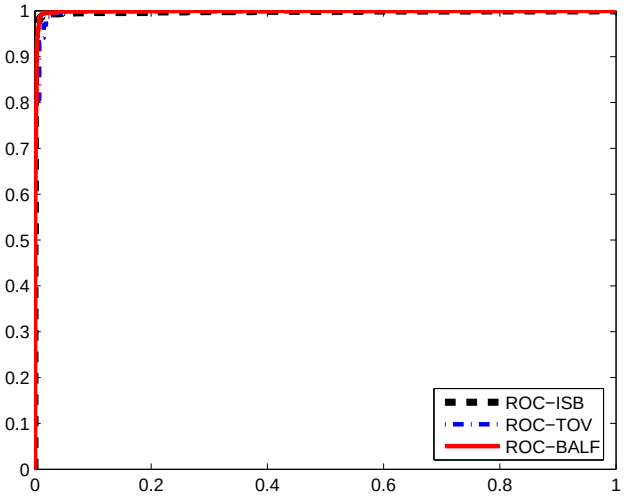


Figure 3.2: ROC of ISB, TOV, and BALF with three most significant features. $\mathrm{AUC_{ISB}} = 0.9976$, $\mathrm{AUC_{TOV}} = 0.9970$, $\mathrm{AUC_{BALF}} = 0.9984$.

Table 3.2: AUC of classifiers built with each feature.

	ISB	TOV	BALF
δ_{cp}	0.9832	0.9839	0.9613
$\delta_{\mathrm{R_{cp}}}$	0.9905	0.9856	0.9964
I_{dc}	0.8973	0.9268	0.8190
n_{bs}	0.6624	0.6476	0.5124

Comparison with existing methods

Since the number of basic sites is not finally determined, we compare the results given in [1] with our results obtained with the other three features, which is shown in Table 3.3. By testing on the same ISB dataset, the proposed features can achieve both higher precisions for doubly and triply charged MS/MS as well as a higher accuracy for all spectra. This indicates that the three features are significant in discriminating doubly charged MS/MS from triply charged ones. Besides, testing these features on the other two independent datasets indeed verify their discriminant power.

Table 3.3: Results obtained by using three features on ISB dataset and the caparison with the results given in [1] on the same dataset are provided.

Features		Estimated Parameters		Precision		Accuracy
		+2	+3	+2	+3	
GMM	δ_{cp}	-0.1175	-1.8433	0.9803	0.9886	0.9833
	$\delta_{\mathrm{R_{cp}}}$	0.8228	-0.8352			
	I_{dc}	0.2847	1.6196			
SVM	see [1]	N/A		0.9240	0.9380	0.9310

Discussion of the number of basic sites

The result about the discriminant power of each feature shows that the number of basic sites is not powerful in discriminating peptides with different charges. The reason is that the computation of this feature is not quite precise. It is hard to compute the number of basic sites, because it is complicated by the following factors: (1) it is possible that the mass differences between many pairs of peaks correspond to one same basic site, because 6 kinds of ions can be generated in CID although they are not equally likely generated. Besides, those ions can produce variants by losing water, ammonia, CO or NH group. (2) When we compute the number of basic sites, we don't want to consider too much about their positions in a sequence, otherwise, it would become another complex problem, peptide de novo sequencing. However, when there are multiple basic sites especially multiple identical basic sites like two K's or R's existing in a peptide, we need to find a way to differentiate these two K's or R's. (3) Situations when tryptic peptides end with two adjacent basic sites (KK, RR, KR, RK, HK, HR) or start with a basic site also complicate the computation. The research in [15] shows that when two basic sites are adjacent, it is more likely that only one of them can attach protons because there exists strong Coulombic repulsion force between adjacent protons. In addition, peptides starting with basic residues will make the N-terminal amine group less likely to attract protons, because the side chains of basic residues have much higher proton affinities than the amine group [15].

According to the definition of n_{bs}, we can approach its computation in two possible ways: (1) compute the pseudo-number of basic sites by counting the number of all cases corresponding to a basic site while ignoring duplicate cases. This is reasonable because the pseudo-number of triply charged peptides should be generally larger than that of doubly charged ones. And (2), figure out a theoretical repeat number of basic sites with the statistics of mass spectrometry generating ions. There is some research conducted to quantify the percentage of each kind of ion produced in CID. The study [16] reports some of such statistics based on the yeast proteome. However, data in a more general sense is needed. With the statistics of ions produced in CID, we can compute a theoretical repeat number for each basic residue. Then, it can be combined with

the pseudo-number to derive the real number of basic sites in a mass spectrum. In this study, the feature n_{bs} was computed as the pseudo-number and transformed to have a variance of 1. This feature is cogent in theory to discriminate doubly and triply charged MS/MS, but how to precisely compute it is still an open problem.

3.4 Conclusions

A new approach for assigning charge states to low-resolution CID MS/MS is proposed based on the unsupervised GMM with four novel and discriminant features extracted from MS/MS. ROC and AUC demonstrate that GMM with proposed features is very promising in classifying doubly and triply charged MS/MS. For future work, we will examine more on the computation of the number of basic sites, which theoretically should be the most significant feature in discriminating peptides with different charges.

Authors contributions

JS developed the algorithm, designed and executed all experimental work, and wrote the first draft. FXW supervised and initiated the project, and revised the manuscript. Both authors read and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors would like to thank Dr. Andrew Keller from Institute for Systems Biology for generously providing spectral data and protein databases for the ISB dataset and Dr. Guy G. Poirier from Laval University for providing the TOV dataset.

References

[1] S. Na, E. Paek, and C. Lee, "Cifter: Automated charge-state determination for peptide tandem mass spectra," *Anal. Chem.*, vol. 80, pp. 1520–1528, 2008.

[2] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, vol. 20, pp. 3551–3567, 1999.

[3] J. K. Eng, A. L. McCormack, and J. R. Y. III, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *J Am Soc Mass Spectrom*, vol. 5, pp. 976–989, 1994.

[4] J. M. Hogan, R. Higdon, N. Kolker, and E. Kolker, "Charge state estimation for tandem mass spectrometry proteomics," *OMICS*, vol. 9, pp. 233–249, 2005.

[5] V. Dancik, T. A. Addona, K. R. Clauser, J. E. Vath, and P. A. Pevzner, "De novo peptide sequencing via tande mass spectrometry," *J. Comput Biol*, vol. 6, pp. 327–342, 1999.

[6] A. A. Klammer, C. C. Wu, M. J. MacCoss, and W. S. Noble, "Peptide charge state determination for low-resolutino tandem mass spectra," in *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference*, 2005.

[7] A. M. Zou, J. Shi, J. Ding, and F. X. Wu, "Charge state determination of peptide tandem mass spectra using support vector machine (svm)," *IEEE Trans Inf Technol B*, vol. 14, pp. 552–558, 2010.

[8] D. L. Tabb, J. K. Eng, and J. R. Yates, "Protein identification by sequest," in *Proteome Research: Mass Spectrometry*, P. James, Ed. Berlin: Springer, 2001, pp. 126–142.

[9] F. X. Wu, P. Gagne, A. Droit, and G. G. Poirier, "Quality assessment of peptide tandem mass spectra," *BMC Bioinformatics*, vol. 9(Suppl 6), p. S13, 2008.

[10] M. Kinter and N. E. Sherman, *Protein sequencing and identification using tandem mass spectrometry.* United States: John Wiley & Sons, Inc, 2000.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning.* United States: Springer, 2006.

[12] ——, *Neural Networks for Pattern Recognition.* United States: Oxford University Press, 1995.

[13] A. Keller, S. Purvine, A. I. Nesvizhskii, S. Stolyar, D. R. Goodlett, and E. Kolker, "Experimental protein mixture for validating tandem mass spectral analysis," *OMICS*, vol. 6, pp. 207–212, 2002.

[14] J. P. Gagne, P. Gagne, J. M. Hunter, M. E. Bonicalzi, J. F. Lemay, I. Kelly, C. L. Page, D. Provencher, A. M. Mes-Masson, A. Droit, D. Bourgais, and G. G. Poirier, "Proteome profiling of human epithelial ovarian cancer cell line tov-112d," *Mol. Cell. Biochem*, vol. 275, pp. 25–55, 2005.

[15] D. L. Tabb, Y. Huang, V. H. Wysocki, and J. R. Yates, "Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of ppetides," *Anal. Chem.*, vol. 76, pp. 1243–1248, 2004.

[16] D. L. Tabb, L. L. Smith, L. A. Breci, V. H. Wysocki, D. Lin, and J. R. Yates, "Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides," *Anal. Chem.*, vol. 75, pp. 1155–1175, 2003.

Chapter 4

Improve accuracy of peptide identification with consistency between peptides

Published as: Jinhong Shi, Bolin Chen and Fang-Xiang Wu, "Improve accuracy of peptide identification with consistency between peptides," *IEEE BIBM'2011*, Atlanta, America, 12-15 November 2011.

We have introduced the processing of MS/MS data, which is in the first computational phase of MS-based protein inference, in the previous chapter. This chapter will address the postprocessing of peptide identification results in the second computational phase. Usually, statistical analysis is performed to postprocess peptide identification results. This step plays a very important role in peptide identification. Today, the amount of MS/MS data is usually very large, and it is impractical to manually verify peptide identification results. As is known, there exists a high rate of false positives in peptide identification. This will bring many false identifications directly to protein inference. Therefore, it is necessary to develop reliable methods to verify the identification results. In addition, statistical analysis of peptide identification facilitates protein inference by providing more reliable measurement of peptide identification accuracy.

The manuscript included in this chapter proposes a new method to estimate the accuracy of peptide identification with logistic regression based on Sequest scores. Instead of using original Sequest scores ΔCn and $Xcorr$, the regularized scores ΔCn^* and $Xcorr^*$ are used as input into the logistic regression model. The scores are regularized by use of an adjacency matrix describing the sibling relationship between peptides.

The results have shown that the proposed method can robustly assign accurate probabilities to peptides and have a very high discrimination power.

Improve accuracy of peptide identification with consistency between peptides

Abstract

A new method is presented to estimate the accuracy of peptide identification with logistic regression (LR) based on Sequest scores. Each peptide is characterized with the regularized Sequest scores ΔCn^* and $Xcorr^*$. The score regularization is formulated as an optimization problem by applying two assumptions: the smoothing consistency between sibling peptides and the fitting consistency between original scores and new scores. An adjacency matrix is built to describe the affinity between peptides, and is used in the score regularization to compute new scores. Then, the new scores are input to the LR model, which is solved with the penalized Newton-Raphson method. By applying the method on two datasets with known validity, the results have shown that the proposed method can robustly assign accurate probabilities to peptides and have a very high discrimination power, higher than that of PeptideProphet, to distinguish correct and incorrect peptides.

4.1 Introduction

Peptide identification by tandem mass spectrometry is an important step in proteomics. One popular way to identify peptides is database searching. The procedure of this method is: First, database proteins are cut to produce peptides in terms of enzyme specificities; Second, theoretical tandem mass spectra (MS/MS) of these peptides are generated; Third, a scoring system is used to measure the similarity between the experimental MS/MS and the theoretical MS/MS, i.e., performing peptide-spectrum-match (PSM); Finally, the peptide with the highest score is usually reported to be identified. Many search engines have been developed for peptide identification and the main difference between them lies in the scoring system they use. This leads to the situation that one query spectrum will have two different sets of scores after being searched with, for instance, Mascot [1] and Sequest [2]. As such, it is hard to compare the search results with

these scores. In addition, peptide identification is not the final goal of proteomics. It lays the foundation of protein inference. To facilitate the comparison between search results and subsequent protein analysis, it is necessary to estimate the accuracy of peptide identification [3–5].

This paper will show a new method of estimating the accuracy of peptide identification based on Sequest search results, though it can be easily and readily extended to other search engines. There have been several methods proposed to improve the accuracy of Sequest peptide identification [3, 6, 7]. The most commonly-used one is PeptideProphet [3]. It used a bimodal and EM algorithm to assign probabilities to Sequest search results. This algorithm has also been extended to analyze search results from X!Tandem [8] and Mascot. The advantage of this algorithm is that it models the distribution of all discriminant scores in a sample, and then uses a Bayesian model to assign probabilities to peptide identifications. However, the probability model heavily depends on the appropriate distribution hypotheses, which needs to be closely verified for each different data set.

In this paper, we propose a new method to assign probabilities to Sequest peptide identifications by use of logistic regression (LR). Rather than inputting the original Sequest scores to the LR model, we first regularize the scores by applying a smoothing consistency assumption between sibling peptides and a fitting consistency assumption between the new scores and original scores. The consistency assumptions have been widely used in semi-supervised learning problems and state that (1) nearby points are likely to have the same label; and (2) points in the same cluster are likely to have the same label [9]. They point out the local and global property of points in different clusters, respectively. As for peptide identification, we can similarly define the nearby peptides as the "sibling peptides", which are generated by the same protein. The smoothing consistency means that sibling peptides should have similar scores. Since peptide identification is preliminarily done with search engines, the original search scores provide the basic clusters of true and false peptides. Thus, the fitting consistency means that the new scores cannot deviate too much from the original scores such that they can keep the basic clusters of peptides.

The consistency assumptions have been applied to peptide identification by He *et al* [10]. In this study, to realize the smoothing consistency, we first propose to use a simpler "peptide-by-peptide" adjacency matrix to replace the "PSM-by-PSM" weight matrix in He's method. The element in the adjacency matrix shows whether two peptides are from one protein or not. He built the weight matrix by presuming that (1) given two peptides from one protein, they are independent of each other; (2) shared peptides are equally generated by parent proteins. The element in the weight matrix denotes the probability that the peptides of two PSMs are from one protein. The results we have got show that the affinity between peptides dominates the smoothing consistency rather than the probability of affinity. Thus, we use a simpler adjacency matrix instead of the weight matrix in the score regularization. Then, we characterize peptides with the regularized scores and build an LR model to assign probabilities to identified peptides.

The Sequest scores ΔCn and $Xcorr$ are used as original scores, and the regularized scores are input to the LR model, which is solved with the penalized Newton-Raphson method. Two datasets are used to evaluate the performance of our method. The results have shown that the assigned probabilities are accurate and have a high power to discriminate correct and incorrect peptide identifications, which is also higher than that of PeptideProphet. Furthermore, we apply the score regularization to PeptideProphet probabilities. It shows that the regularized results have a higher discrimination power than PeptideProphet as well.

4.2 Methods and materials

4.2.1 Workflow of peptide identification

The goal of peptide identification is to obtain peptides along with their probabilities by interpreting MS/MS data. As is known, one peptide could be matched by different MS/MS. Although the number of MS/MS and their matched scores can reflect the confidence of an identified peptide, they are only interpreted to the

same peptide. In other words, the interpreting result from a group of PSMs corresponding to one peptide is equivalent to the interpreting result from the best PSM in this group. Here, we do not take the goal of peptide identification as to interpret each MS/MS, i.e., PSM. Instead, we set our goal as obtaining identified peptides and their probabilities, which can be directly used by protein inference. Therefore, we perform peptide identification based on Sequest with the following procedures.

- PSM: use Sequest to perform PSM and get the original scores of peptide identification.

- Filtering: use PeptideProphet to filter peptide identifications. This step can be easily included since PeptideProphet is in the Trans-Proteomics-Pipeline (TPP) [11], and is free to users and it can analyze the results from commonly-used search engines including Sequest, Mascot, X!Tandem and so on. It has been shown that $80 - 90\%$ of reported peptide identifications are incorrect if the results are not filtered [12]. Thus, it is reasonable to only use the filtered results (default setting: probability≥ 0.05) from PeptideProphet for the analysis. Moreover, it can also save time and resources to handle a much smaller amount of data.

- Build peptide-protein relation matrix W_0: the element in the matrix indicates whether a peptide belongs to a protein or not.

- Build peptide-peptide adjacency matrix W: the element in the matrix indicates whether two peptides are from one protein or not. In order to cancel the self-enforcement effects, the diagonal elements are set to zeros.

- Perform score regularization with the adjacency matrix.

- Compute probabilities of peptide identification using logistic regression with regularized scores.

- Output identified peptides and their probabilities by setting proper thresholds.

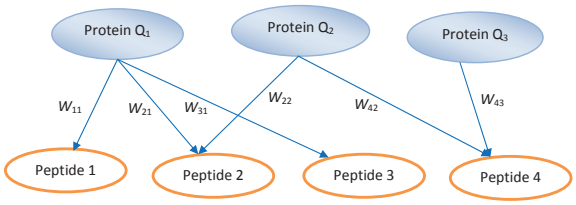


Figure 4.1: A configuration to show the construction of peptide-protein relation matrix W_0.

4.2.2 Score regularization

This section will first introduce the construction of adjacency matrix and then describe the regularization of search scores.

Construction of adjacency matrix

According to the workflow of peptide identification, suppose that L PSMs passed the default filtering of PeptideProphet, and they correspond to N peptides and M proteins. First, we will use these N peptides and M proteins to construct an N by M peptide-protein relation matrix W_0. The element in the matrix shows the belonging relationship of a peptide to a protein. A simple example is given to show the construction of peptide-protein relation matrix W_0 in Figure 4.1.

The relation matrix W_0 is given as

$$W_0 = \begin{bmatrix} w_{0_{11}} & 0 & 0 \\ w_{0_{21}} & w_{0_{22}} & 0 \\ w_{0_{31}} & 0 & 0 \\ 0 & w_{0_{42}} & w_{0_{43}} \end{bmatrix},$$

where $w_{0_{ij}}$ denotes the relationship between peptide i and protein j. If a peptide belongs to a protein, then $w_{0_{ij}} = 1$; otherwise, $w_{0_{ij}} = 0$. Then, we construct a matrix between peptides as follows,

$$W' = W_0 * W_0^T, \tag{4.1}$$

where W_0^T is the transpose of W_0. In W', $w_{ij} = 0$ if peptide i and peptide j are not from any protein; $w_{ij} = 1$ if two peptides are only from one protein. It is possible that two peptides are simultaneously shared by multiple proteins, in this case, $w_{ij} > 1$. However, we only consider the notion that two peptides are siblings or not, thus we set all $w_{ij} = 1$ as long as peptide i and peptide j are siblings, no matter how many parent proteins they may have. In addition, in order to cancel the self-enforcement effect of peptides, the diagonal elements of W' are set zeros, i.e., $w_{ii} = 0$. We call this newly derived matrix as the adjacency matrix between peptides, and denote it as W in the following.

Here, W is a symmetrical matrix. Thus, the confidence enforcement is spread symmetrically among sibling peptides. In essence, the adjacency matrix accounts for extra information from proteins, i.e., using proteins to build the affinities between peptides, which is the reason why the later score regularization can improve the discrimination power of scores.

Here, we also build a diagonal matrix D from W. The diagonal elements are defined as $d_{ii} = \sum_{k=1}^{N} w_{ki}$. The diagonal matrix D is called the degree matrix in spectral graph theory [13]. The differences between the adjacency matrix and He's weight matrix are summarized as follows: We build the adjacency matrix with the goal of identifying peptides, rather than interpreting all MS/MS data, i.e., PSM. Thus, our adjacency matrix is of (the number of) peptides by (the number of) peptides. Instead, He's matrix is of (the number of) PSM by (the number of) PSM. Compared to He's method, our adjacency matrix is easier to build, much smaller and more goal-driven in peptide identification. However, if the goal is to study the interpretability of MS/MS from a certain instrument, we need to build the PSM by PSM matrix.

Score regularization

As our goal is to get peptides that can be interpreted from MS/MS, for each peptide, we only choose the PSM with the highest PeptideProphet probability as the evidence to show its identification. The original scores of these N peptides are given by $X = (x_1, x_2, ..., x_N)$, selected from the L PSM. The x_i could be a

scalar, such as the negative logarithm of E-value from X!Tandem, or a vector, such as Sequest ($\Delta Cn, Xcorr$).

Given the vector of original scores X, and the adjacency matrix W, we compute a vector of new scores Y by simultaneously applying the consistency assumptions: smoothing consistency among sibling peptides and fitting consistency between original scores and new scores [9, 10].

Smoothing consistency: The inconsistency of scores of sibling peptides is formulated in the following cost function,

$$S(Y) = \frac{1}{2} \sum_{i,j=1}^{N} w_{ij} \left(\frac{y_i}{\sqrt{d_{ii}}} - \frac{y_j}{\sqrt{d_{jj}}} \right)^2, \tag{4.2}$$

where d_{ii} and d_{jj} are the diagonal elements of the degree matrix D. If sibling peptides have quite different scores, then the cost function value will be large. Here, we can see that d_{ii} cannot be *zero*. From the definition of D, $d_{ii} = 0$ means that the peptide i has no siblings. In order to avoid this situation, we add a dummy sibling to such peptides by setting d_{ii} to be a fairly small value as in [10]. In addition, we write $S(Y)$ in a matrix form as follows,

$$S(Y) = Y^T (I - D^{-1/2} W D^{-1/2}) Y, \tag{4.3}$$

where I is the identity matrix. The derivation of Equation(4.3) from Equation(4.2) can be found in [10].

Fitting consistency: The inconsistency between original scores and new scores is given by

$$F(Y) = \sum_{i=1}^{N} (y_i - x_i)^2. \tag{4.4}$$

This value will be large if the new scores deviate too much from the original scores.

Objective function: A linear combination of $S(Y)$ and $F(Y)$ is used to compose the final cost function,

$$Q(Y) = (1 - \lambda) S(Y) + \lambda F(Y), \tag{4.5}$$

where $\lambda \in (0, 1)$ is the parameter which can regularize the balance between the smoothing consistency and the fitting consistency. Then, the objective becomes to find the new scores Y^* which can minimize $Q(Y)$, i.e.,

$$Y^* = \arg\min_Y Q(Y). \tag{4.6}$$

By taking the derivative of $Q(Y)$ w.r.t. Y, and setting the derivative zero, we get

$$Y^* = \lambda(I - (1-\lambda)V)^{-1}X, \tag{4.7}$$

where $V = D^{-1/2}WD^{-1/2}$.

4.2.3 Logistic regression

LR is used to represent the posterior probabilities of peptide identifications given the scores. Under the general assumptions [14], the posterior probability of a random peptide identification Z can be written as a sigmoid function acting on a linear combination of a feature vector $\boldsymbol{\phi}$ so that

$$p\left(Z=1\middle|\boldsymbol{\phi},\mathbf{w}\right) = \frac{\exp\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\right)}{1+\exp\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\right)} \tag{4.8}$$

$$p\left(Z=0\middle|\boldsymbol{\phi},\mathbf{w}\right) = \frac{1}{1+\exp\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\right)}, \tag{4.9}$$

where $\mathbf{w}$ is a weight vector. Here, $p\left(Z=1\middle|\boldsymbol{\phi},\mathbf{w}\right)$ and $p\left(Z=0\middle|\boldsymbol{\phi},\mathbf{w}\right)$ are the posterior probabilities of a correct and incorrect peptide identification given its feature vector $\boldsymbol{\phi}$, respectively.

We build the feature vector as $\boldsymbol{\phi} = (\Delta Cn^*, Xcorr^*)$. Different feature vectors can be built for other search engines [3, 5]. The weight vector $\mathbf{w}$ is estimated directly from the new scores and the given labels of peptides by maximizing the conditional likelihood, i.e.,

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \left(\prod_{i=1}^{N} p\left(Z^i\middle|\boldsymbol{\phi}^i,\mathbf{w}\right) \right), \tag{4.10}$$

where N is the number of peptides, Z^i is the i^{th} peptide, and $\boldsymbol{\phi}^i$ is its feature vector. Equation (4.10) is equivalent to maximizing the conditional log likelihood

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \mathcal{L}(\mathbf{w}), \tag{4.11}$$

where the conditional log likelihood $\mathcal{L}(\mathbf{w})$, after substitutions of (4.8) and (4.9) into (4.10) and some mathematical manipulations, is given by

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} \left[Z^i \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}^i \; - \; \ln\left(1 + \exp\left(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}^i\right)\right)\right]. \tag{4.12}$$

However, there is no analytic solution to (4.11). Thus, we choose the Newton-Raphson method to find a numerical solution. To avoid over-fitting, a penalty $\frac{\lambda_1}{2} \left\|\mathbf{w}\right\|_2^2$ is imposed on large fluctuations of the parameter $\mathbf{w}$. In addition, we also add a prior knowledge μ_0 to the regularization term. The penalized log likelihood function $\mathcal{L}_P(\mathbf{w})$ now is written as

$$\mathcal{L}_P(\mathbf{w}) = \mathcal{L}(\mathbf{w}) - \frac{\lambda_1}{2} \left\|\mathbf{w} - \mu_0\right\|_2^2, \tag{4.13}$$

where λ_1 is a constant determining the strength of the penalty term. By taking the first and second derivatives of $\mathcal{L}_P(\mathbf{w})$ w.r.t. the weight vector $\mathbf{w}$, we have the iterative equation of the Newton-Raphson method as follows,

$$\hat{\mathbf{w}}_{i+1} = \hat{\mathbf{w}}_i - \left[\nabla^2 \mathcal{L}\left(\hat{\mathbf{w}}_i\right) - \lambda_1 I\right]^{-1} \left[\nabla \mathcal{L}(\hat{\mathbf{w}}_i) - \lambda_1 (\hat{\mathbf{w}}_i - \mu_0)\right] \tag{4.14}$$

where $\nabla^2 \mathcal{L}\left(\hat{\mathbf{w}}_i\right)$ is the Hessian matrix, and $\nabla \mathcal{L}(\hat{\mathbf{w}}_i)$ is the scoring function of $\mathcal{L}(\mathbf{w})$ at the i^{th} iteration. Since $\mathcal{L}_P(\mathbf{w})$ is concave w.r.t. $\mathbf{w}$, because the Hessian matrix $\nabla^2 \mathcal{L}\left(\hat{\mathbf{w}}_i\right)$ is semi-negative definite, this method will always converge to a global maximum. An active learning method is used to select training data for solving the LR model. The details can be found in a previous work [5].

4.2.4 Experimental Data

Two datasets are downloaded from ISB public database and the detailed description of the data can be found in reference [15]. These two datasets are generated from two mixtures in which there are 18 standard proteins and 15 contaminants also considered to be present. One dataset is generated on the Thermo Electron (Waltham, MA) LTQ, called Mix1_LTQ. The other one is produced with ABI (Foster City, CA) API QSTAR Pulsar i, called Mix2_QSTAR. The datasets from different instruments are used to verify the robustness of

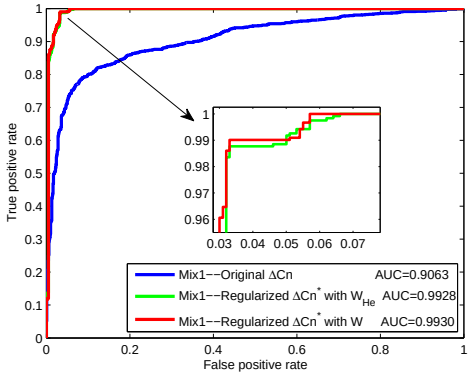


Figure 4.2: The ROC of the original ΔCn, the regularized ΔCn^* with our matrix W and W_{He} for Mix1.

our method to datasets from various experiments. The statistics of the two datasets is summarized in Table 4.1.

Table 4.1: Statistics of the two datasets: the number of MS/MS, the number of PSM passed PeptideProphet default filtering, the number of peptides and proteins corresponding to these PSM.

	MS/MS	PSM passed the filter	Peptides	Proteins
Mix1	86850	19814	2217	729
Mix2	26780	6929	1200	383

4.3 Results

Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) are employed to measure the discrimination power of the regularized scores and LR model. ROC curves of actual classifications locate in between the ideal plot (the point $(0, 1)$) and the random-guess plot (the diagonal line) with AUC $\in (0.5, 1)$. The bigger the AUC, the higher the discrimination power.

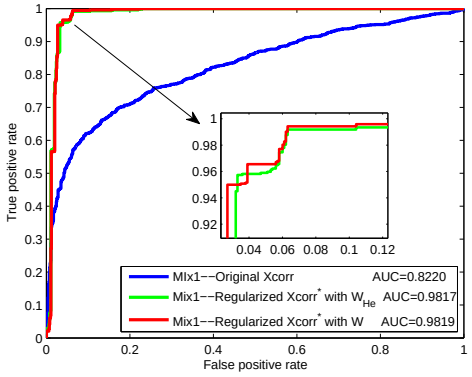


Figure 4.3: The ROC of the original $Xcorr$, the regularized $Xcorr^*$ with our matrix W and W_{He} for Mix1.

4.3.1 Regularized Sequest scores

We first demonstrate the results of applying our method on Sequest scores ΔCn and $Xcorr$, which are given in Figure 4.2 to Figure 4.5. They show that

- The regularized scores, ΔCn^* and $Xcorr^*$, can greatly improve the discrimination power between correct and incorrect peptide identifications.

- The discrimination power of regularized scores are consistent with the original scores. We can see that Sequest ΔCn has a higher discrimination power than $Xcorr$, so does the regularized ΔCn^*.

- The discrimination power of regularized scores heavily depends on the original scores. This is verified by the difference between He's results and our results on the same dataset Mix2_QSTAR. He used X!Tandem negative logarithm of E-value as the original scores [10]. The AUC of X!Tandem original scores is 0.64, while the regularized scores can only marginally improve the AUC to 0.65. However, by using Sequest scores of the same dataset, the regularized scores can significantly improve the AUC from $A_{\Delta Cn} = 0.8626$ and $A_{Xcorr} = 0.8011$ to 0.9882 and 0.9871, respectively.

- There is a very small difference between the discrimination power of the scores regularized with the

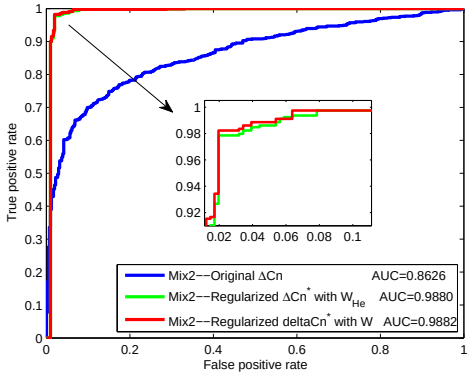


Figure 4.4: The ROC of the original ΔCn, the regularized ΔCn^* with our matrix W and W_{He} for Mix2.

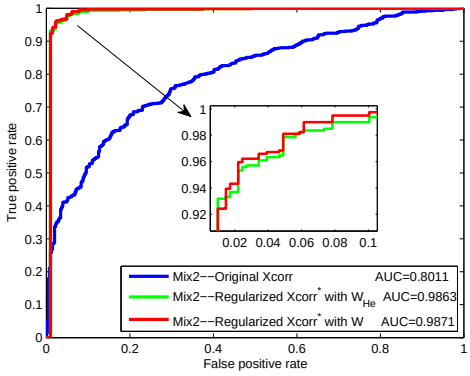


Figure 4.5: The ROC of the original $Xcorr$, the regularized $Xcorr^*$ with our matrix W and W_{He} for Mix2.

proposed simpler adjacency matrix and He's matrix. The proposed adjacency matrix can produce a slightly higher power than He's matrix, see the red line in magnified subfigures. This implies that the affinity between sibling peptides dominates the regularization, while the effect of the probability of affinity is quite small.

- The score regularization is robust to different datasets. The results for Mix1 and Mix2 have the same trends and they are both of very good performance.

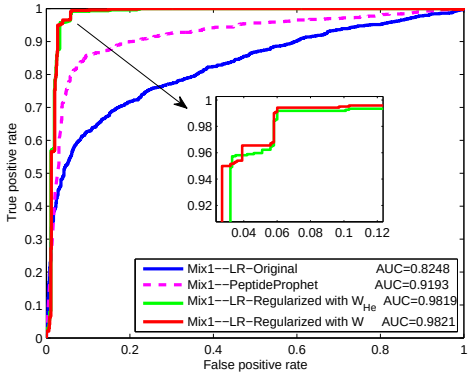


Figure 4.6: ROC of logistic regression based on original and regularized Sequest scores, as well as the ROC of PeptideProphet results for Mix1.

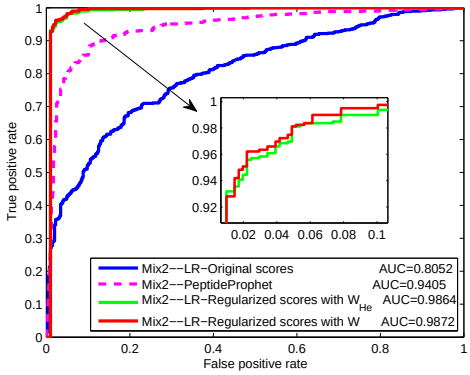


Figure 4.7: ROC of logistic regression based on original and regularized Sequest scores, as well as the ROC of PeptideProphet results for Mix2.

4.3.2 Logistic regression results

We compute the logistic regression with the original and regularized Sequest scores ΔCn and $Xcorr$. The results are illustrated in Figure 4.6 and Figure 4.7. It can be seen that:

- The discrimination power of the LR results based on the Sequest original scores is the lowest one among the four situations. The discrimination power of PeptideProphet is much lower than that of the LR results computed from the regularized scores. As is known, PeptideProphet is the most commonly-used program for the accuracy estimation of peptide identification. It incorporates Sequest

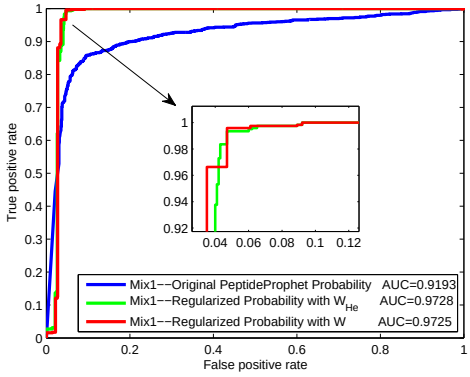


Figure 4.8: ROC of original and regularized PeptideProphet probabilities for Mix1.

ΔCn, $Xcorr$, $SpRank$ and d_M to get discriminant scores, and then applies EM algorithm to estimate the probabilities of peptide identification. Compared to this method, the probabilities generated by logistic regression based on regularized Sequest ΔCn^* and $Xcorr^*$ are more accurate and easier to get.

- Again, the discrimination power of LR model computed with scores regularized with He's matrix and the proposed simpler adjacency matrix is very similar.

4.3.3 Regularized PeptideProphet probabilities

We also apply our method on PeptideProphet probabilities. Figure 4.8 and Figure 4.9 show that the regularized probabilities have an obvious improved discrimination power over the original PeptideProphet probabilities. Since PeptideProphet already accounts for the information of each individual peptide, by including the affinity information between peptides, the regularization actually improves the scores by spreading the confidence of peptides to their siblings.

At very low false positive rates (FPR < 0.03) for Mix1 (see Figure 4.8), the regularized probabilities give a lower true positive rates (TPR) than PeptideProphet. However, when analyzing peptide identification results, the FPR usually takes the value of 0.05. In this case, the regularized probabilities can yield a much higher TPR than PeptideProphet.

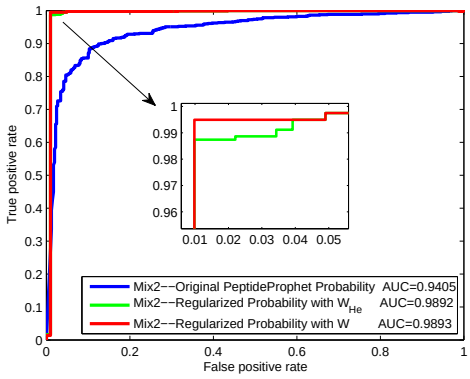


Figure 4.9: ROC of original and regularized PeptideProphet probabilities for Mix2.

4.4 Conclusion

We have demonstrated a new method to assign probabilities to identified peptides. By testing our method on two datasets, the results have shown that the new method can robustly assign accurate probabilities to identified peptides and indeed have a very high power to distinguish correct and incorrect peptides. Furthermore, the discrimination power of the new method is also higher than that of PeptideProphet, the most commonly-used program to assign probabilities to peptide identifications. Compared to PeptideProphet, in addition to the higher discrimination power, the new method also has some other benefits: it is easier, faster, and more goal-driven in peptide identification. The peptides and their probabilities output from the method can be directly used for the subsequent protein inference. Furthermore, it is robust to different experimental datasets. However, the distribution hypotheses in PeptideProphet need careful examinations when different datasets are used. Although we exemplify the procedure with Sequest search results, the method can be easily and readily extend to other search engines.

Acknowledgment

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] D. N. Perkins, D. J. C. Pappin, D. M. Creasy and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, 20: 3551-3567, 1999.

[2] J. K. Eng, A. L. McCormack and J. R. Yates III, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *J. Am. Soc. Mass Spectrom.*, 5:976-989, 1994.

[3] A. Keller, A. I. Nesvizhskii, E. Kolker and R. Aebersold, "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search," *Anal. Chem.*, 74: 5383-5392, 2002.

[4] A. I. Nesvizhskii, A. Keller, E. Kolker and R. Aebersold, "A statistical model for identifying proteins by tandem mass spectrometry," *Anal. Chem.*, 75: 4646-4658, 2003.

[5] J. Shi, W. Lin, and F.-X. Wu, "Statistical analysis of Mascot peptide identification with active logistic regression," *iCBBE*, 2010.

[6] C. Shao, W. Sun, F. Li, R. Yang, L. Zhang and Y. Gao, "OScore: a combined score to reduce false negative rates for peptide identification in tandem mass spectrometry analysis," *J. Mass. Spectrom.*, 44: 25-31, 2009.

[7] J. Razumovskaya, V. Olman, D. Xu, E. C. Uberbacher, N. C. VerBerkmoes, R. L. Hettich, and Y. Xu, "A computational method for assessing peptide identification reliability in tandem mass spectrometry analysis with Sequest," *Proteomics*, 4: 961-969, 2004.

[8] R. Craig, and R. C. Beavis, "TANDEM: matching proteins with tandem mass spectra," *Bioinformatics*, 20: 1466-1467, 2004.

[9] D. Zhou, O. Bousquet, T. N. Lal, J. Weston and B. Schölkopf, "Learning with local and global consistency." *NIPS*, 2003.

[10] Z. He, H. Zhao and W. Yu, "Score regularization for peptide identification," *BMC Bioinfomatics*, 12(Suppl):S2, 2011.

[11] A. Keller, J. Eng, N. Zhang, X.-J. Li and R. Aebersold, "A uniform proteomics MS/MS analysis platform utilizing open XML file formats," *Mol. Syst. Biol.*, 2005.

[12] A. Keller, S. Purvine, A. I. Nesvizhskii, S. Stolyar, D. R. Goodlett and E. Kolker, "Experimental protein mixture for validating tandem mass spectral analysis," *OMICS*, 6(2): 207-212, 2002.

[13] D. Spielman. Spectral Graph Theory, *in* Combinatorial Scientific Computing, Chapman and Hall/CRC Press. 2010.

[14] C. M. Bishop. Pattern Recognition and Machine Learning. Springer. Singapore, 2006.

[15] J. Klimek, J. S. Eddes, L. Hohmann, J. Jackson, A. Peterson, S. Letarte, P. R. Gafken, J. E. Katz, P. Mallick, H. Lee, A. Schmidt, R. Ossola, J. K. Eng, R. Aebersold and D. B. Martin, "The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools," *J. Proteome Res.*, 7: 96-103, 2008.

Chapter 5

Unifying protein inference and peptide identification with feedback to update consistency between peptides

Published as: Jinhong Shi, Bolin Chen and Fang-Xiang Wu, "Unifying protein inference and peptide identification with feedback to update consistency between peptides," *Proteomics*, accepted.

To this point, we have introduced the processing of MS/MS data and the postprocessing of peptide identification results in the last two chapters. This chapter will discuss protein inference with the identified peptides and their probabilities. Existing methods to address this problem can be classified into two groups: two-stage and one unified framework to perform peptide identification and protein inference. In two-stage methods, protein inference is based on, but also separated from, peptide identification. While in one unified framework, protein inference and peptide identification are integrated together by adding a feedback from protein inference to peptide identification. This feedback can improve peptide identification results, and in turn increase the accuracy and coverage of inferred proteins.

The manuscript included in this chapter proposes an iterative method to infer proteins based on peptides identified from Sequest. The statistical analysis of peptide identification is performed with the logistic regression that has been introduced in the previous chapter. Protein inference and peptide identification are iterated in one framework by adding a feedback from protein inference to peptide identification. The

feedback information is a list of high-confidence proteins, which is used to update the adjacency matrix between peptides. The adjacency matrix is then used in the regularization of peptide scores. The results have shown that the proposed method can infer more true positive proteins, while outputting less false positive proteins than ProteinProphet [1] at the same false positive rate. The coverage of inferred proteins is also significantly increased due to the selection of multiple peptides for each MS/MS and the improvement of their scores by the feedback from the inferred proteins.

Unifying protein inference and peptide identification with feedback to update consistency between peptides

Abstract

We first propose a new method to process peptide identification reports from databases search engines. Based on it, we then develop a method for unifying protein inference and peptide identification by adding a feedback from protein inference to peptide identification. The feedback information is a list of high-confidence proteins, which is used to update an adjacency matrix between peptides. The adjacency matrix is used in the regularization of peptide scores. Logistic regression (LR) is used to compute the probability of peptide identification with the regularized scores. Protein scores are then calculated with the LR probability of peptides. Instead of selecting the best peptide match for each MS/MS, we select multiple peptides. By testing on two datasets, the results show that the proposed method can robustly assign accurate probabilities to peptides, and has a higher discrimination power than PeptideProphet to distinguish correct and incorrect identified peptides. Additionally, not only can our method infer more true positive proteins, but also infer less false positive proteins than ProteinProphet at the same false positive rate. The coverage of inferred proteins is also significantly increased due to the selection of multiple peptides for each MS/MS and the improvement of their scores by the feedback from the inferred proteins.

5.1 Introduction

Protein inference by assembling peptides identified from tandem mass spectra is an important computational step in proteomics, based on which further analysis, such as inference of protein structure and function can be performed [2, 3]. This problem has been systematically discussed in [4–6]. Existing methods to address this problem can be split into two groups. The first group performs protein inference and peptide identification separately [1, 7–9]. First, peptides are identified from tandem mass spectra by de novo sequencing [10–

12] or by database searching [13–15]. Then, proteins are inferred by assembling these identified peptides. The other group combines protein inference with peptide identification, identifying peptides and proteins simultaneously [16–18].

Spivak *et al* have built a Barista model [17] which formulates the protein identification as an optimization problem. The protein inference problem is represented as a tripartite graph, with layers corresponding to spectra, peptides and proteins. The input to Barista is the tripartite graph with a set of features that describes matches between peptides and spectra (PSM). The parameters in the model are estimated by training the model with reference data, and then the trained model is used to infer proteins. The advantage of this model is that it utilizes the spectrum information in all the steps of protein inference, without discarding spectra from peptide identification to protein inference. The application of this method is limited by the necessity of reference data to train the model each time when different datasets are analyzed. Since many well-developed search engines for peptide identification are available, methods for processing the peptide identification reports from these engines have been proposed. For example, Li *etc* have used a nested mixture model [18] to estimate peptide and protein probability at the same time based on identified peptides and their scores from search engines. This model allows evidence feedback between proteins and their constituent peptides. It is built on several reasonable assumptions except that it ignores the problem of shared peptides.

This paper proposes a method to unify protein inference and peptide identification by adding a feedback from protein inference to peptide identification. The feedback is applied by use of the smoothing consistency between peptides, which is constructed from the mapping relationship between the inferred proteins and the identified peptides. Similar to [18], we rely the protein inference process on the peptide identification reports from database search engines. However, we select multiple peptides instead of only choosing the best match for each MS/MS. First, we expect that the feedback from protein inference can improve the peptide identification scores, especially of those that are not the best matches. Second, we also expect to improve the coverage of proteins by increasing the number of identified peptides. Two datasets have been used to verify our proposed method, and the results have shown that this feedback method can significantly increase

the number of identified peptides and the coverage of inferred proteins compared to PeptideProphet [19] and ProteinProphet [1], respectively.

5.2 Methods and materials

5.2.1 Feedback workflow for peptide identification and protein inference

The feedback workflow for peptide identification and protein inference is shown in Figure 5.1. The starting point of this workflow are the peptide identification reports from database search engines. In this study, we test our method based on Sequest [14] peptide identification results.

First, multiple peptides are selected for each MS/MS spectrum from Sequest *.out* files; here, we select 3 peptides for each MS/MS. There has been some work to rerank peptide identification results for MS/MS [20], which uses a machine learning method to recompute the coefficients for PeptideProphet [19] model. However, we don't aim to rerank peptide identification results for each MS/MS, but instead, we aim to improve the results with feedback from protein inference. Second, putative peptides are used to search proteins in the database. Third, an adjacency matrix which shows whether two peptides are siblings or not is built according to the list of proteins. Then, peptide scores are regularized with the application of two consistency assumptions, and the regularized scores are used as features in logistic regression (LR) to compute peptide identification probability. Based on the LR probability, protein scores are computed. Next, high-confidence proteins are selected to compose the new list of proteins, which is used to update the adjacency matrix between peptides. The experiments have shown that the loop will stop in two to four iterations for the used datasets.

The main advantage of this workflow is that many peptides that are not selected by, such as PeptideProphet [19], are given the chance to be identified with the help of the feedback from protein inference. In

return, this will significantly improve the coverage of inferred proteins. The following sections will introduce the logistic regression, score regularization and protein inference model, respectively.

5.2.2 Logistic regression to compute peptide identification probability

Logistic regression is used to represent the posterior probabilities of peptide identifications given their scores. Under the general assumptions [21], the posterior probability of a random peptide identification Z can be written as a sigmoid function of a linear combination of a feature vector $\boldsymbol{\phi}$ so that

$$p\left(Z=1\middle|\boldsymbol{\phi},\mathbf{w}\right)=\frac{\exp\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\right)}{1+\exp\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\right)} \tag{5.1}$$

and

$$p\left(Z=0\middle|\boldsymbol{\phi},\mathbf{w}\right)=\frac{1}{1+\exp\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\right)}, \tag{5.2}$$

where $\mathbf{w}$ is a weight vector. Here, $p\left(Z=1\middle|\boldsymbol{\phi},\mathbf{w}\right)$ and $p\left(Z=0\middle|\boldsymbol{\phi},\mathbf{w}\right)$ are the posterior probabilities of a correct and incorrect peptide identification given its feature vector $\boldsymbol{\phi}$, respectively. Notice that Equation(5.2) follows directly from Equation(5.1) because the sum of these two probabilities must be 1.

We build the feature vector as $\boldsymbol{\phi}=(\Delta Cn^{*},Xcorr^{*})$. The notations ΔCn^{*} and $Xcorr^{*}$ represent the regularized Sequest scores, the computation of which will be introduced later. Different feature vectors can be employed for other search engines [19, 22]. The weight vector $\mathbf{w}$ is estimated directly from the scores and the given labels of peptides by maximizing the conditional likelihood, i.e.,

$$\hat{\mathbf{w}}=\arg\max_{\mathbf{w}}\left(\prod_{i=1}^{N}p\left(Z^{i}\middle|\boldsymbol{\phi}^{i},\mathbf{w}\right)\right), \tag{5.3}$$

where N is the number of peptides, Z^{i} is the i^{th} peptide identification, and $\boldsymbol{\phi}^{i}$ is its feature vector. Equation (5.3) is equivalent to maximizing the conditional log likelihood

$$\hat{\mathbf{w}}=\arg\max_{\mathbf{w}}\mathcal{L}(\mathbf{w}), \tag{5.4}$$

where the conditional log likelihood $\mathcal{L}(\mathbf{w})$, after substitutions of Equation(5.1) and (5.2) into Equation(5.3) and some mathematical manipulations, is given by

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} \left[z^i \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}^i \; - \; \ln\left(1 + \exp\left(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}^i\right)\right)\right], \tag{5.5}$$

where $z^i \in \{0, 1\}$ is the label of the i^{th} peptide. However, there is no analytic solution to Equation(5.4). Thus, we choose the Newton-Raphson method to find a numerical solution. To avoid over-fitting to the training data, a penalty $\frac{\lambda_1}{2} \|\mathbf{w}\|_2^2$ is imposed on large fluctuations of the parameter $\mathbf{w}$. The penalized log likelihood function $\mathcal{L}_P(\mathbf{w})$ is written as

$$\mathcal{L}_P(\mathbf{w}) = \mathcal{L}(\mathbf{w}) - \frac{\lambda_1}{2} \|\mathbf{w}\|_2^2, \tag{5.6}$$

where λ_1 is a constant determining the strength of the penalty term. We also add a prior knowledge μ_0 to the regularization term, and the penalized log likelihood function now becomes

$$\mathcal{L}_P(\mathbf{w}) = \mathcal{L}(\mathbf{w}) - \frac{\lambda_1}{2} \|\mathbf{w} - \mu_0\|_2^2. \tag{5.7}$$

By taking the first and second derivatives of $\mathcal{L}_P(\mathbf{w})$ w.r.t. the weight vector $\mathbf{w}$, we have the iterative equation of the Newton-Raphson method as follows,

$$\hat{\mathbf{w}}_{i+1} = \hat{\mathbf{w}}_i - \left[\nabla^2 \mathcal{L}\left(\hat{\mathbf{w}}_i\right) - \lambda_1 I\right]^{-1} \left[\nabla \mathcal{L}(\hat{\mathbf{w}}_i) - \lambda_1 (\hat{\mathbf{w}}_i - \mu_0)\right] \tag{5.8}$$

where $\nabla^2 \mathcal{L}\left(\hat{\mathbf{w}}_i\right)$ is the Hessian matrix, and $\nabla \mathcal{L}(\hat{\mathbf{w}}_i)$ is the scoring function of $\mathcal{L}(\mathbf{w})$ at the i^{th} iteration. Since the penalized conditional log likelihood function $\mathcal{L}_P(\mathbf{w})$ is concave w.r.t. $\mathbf{w}$, because the Hessian matrix $\nabla^2 \mathcal{L}\left(\hat{\mathbf{w}}_i\right)$ is semi-negative definite, this method will always converge to a global maximum [23].

5.2.3 Regularization of peptide scores

This section will first introduce the construction of the adjacency matrix and then describe the regularization of search scores.

Construction of adjacency matrix

Suppose that N peptides are selected and they correspond to M proteins. First, we will use these N peptides and M proteins to construct an N-by-M peptide-protein relation matrix W_0. The element $w_{0_{ij}}$ in the matrix denotes the relationship between peptide i and protein j. If peptide i belongs to protein j, then $w_{0_{ij}} = 1$; otherwise, $w_{0_{ij}} = 0$. Given W_0, we construct a matrix between peptides as follows,

$$W' = W_0 * W_0^T, \tag{5.9}$$

where W_0^T is the transpose of W_0. In W', $w_{ij} = 0$ if peptide i and peptide j are not from any same protein; $w_{ij} = 1$ if two peptides are only from one protein. It is possible that two peptides are simultaneously shared by multiple proteins, in this case, $w_{ij} > 1$. However, we only consider the fact that two peptides are siblings or not, thus we set all $w_{ij} = 1$ as long as peptide i and peptide j are siblings, no matter how many common parent proteins they may have. In addition, in order to cancel the self-enforcement effect of peptides, the diagonal elements of W' are set zeros, i.e., $w_{ii} = 0$. We call this newly derived matrix as the adjacency matrix between peptides, and denote it as W in the following.

Here, W is a symmetrical matrix. Thus, the confidence enforcement is spread symmetrically among sibling peptides. In essence, the adjacency matrix accounts for extra information from proteins, i.e., using proteins to build the affinities between peptides, which is the reason why the regularized scores can improve the discrimination power of LR probability. In addition, we also build a diagonal matrix D from the adjacency matrix W. The diagonal elements are defined as $d_{ii} = \sum_{k=1}^{N} w_{ki}$.

Score regularization

As mentioned before, multiple peptides are selected for each MS/MS. Each peptide may have multiple MS/MS mapped to it, in this case, we find the best MS/MS for each peptide according to the Sequest correlation score $Xcorr$. Then, the two kinds of original scores of the selected N peptides are given by

$X = (x_1, x_2, ..., x_N)$, and x_i is ΔCn_i or $Xcorr_i$. Given the original score X and the adjacency matrix W, we compute a corresponding new score Y by applying the consistency assumptions: smoothing consistency among sibling peptides and fitting consistency between original scores and new scores [24, 25].

Smoothing consistency: the inconsistency between scores of sibling peptides is formulated as,

$$S(Y) = \frac{1}{2} \sum_{i,j=1}^{N} w_{ij} \left(\frac{y_i}{\sqrt{d_{ii}}} - \frac{y_j}{\sqrt{d_{jj}}} \right)^2, \tag{5.10}$$

where d_{ii} and d_{jj} are the diagonal elements of the degree matrix D. If sibling peptides have quite different scores, then the cost function value will be large. Here, we can see that d_{ii} cannot be *zero*. From the definition of D, $d_{ii} = 0$ means that the peptide i has no siblings. In order to avoid this situation, we add a dummy sibling to such peptides by setting d_{ii} to be a fairly small value as in [24]. In addition, we write $S(Y)$ in a matrix form as follows

$$S(Y) = Y^T (I - D^{-1/2} W D^{-1/2}) Y, \tag{5.11}$$

where I is the identity matrix. Note that Y is a column vector here, so $S(Y)$ in the matrix form is still a scalar. The derivation of Equation(5.11) from Equation(5.10) can be found in [24].

Fitting consistency: the inconsistency between original scores and new scores is given by

$$F(Y) = \sum_{i=1}^{N} (y_i - x_i)^2. \tag{5.12}$$

This value will be large if the new scores deviate too much from the original scores.

Objective function: a linear combination of $S(Y)$ and $F(Y)$ is used to compose the final cost function,

$$Q(Y) = (1 - \lambda) S(Y) + \lambda F(Y), \tag{5.13}$$

where $\lambda \in (0, 1)$ is a parameter which can regularize the balance between the smoothing consistency and the fitting consistency. It can be seen that when $\lambda = 1$, the new score equals the original score. Here, we want to regularize the scores more with the smoothing consistency between peptides, thus λ is preferred to take

small values. In this study, we take $\lambda = 0.01$. Then, the objective becomes to find the new scores Y^* which can minimize $Q(Y)$, i.e.,

$$Y^* = \arg\min_Y Q(Y). \tag{5.14}$$

By taking the derivative of $Q(Y)$ w.r.t. Y, and setting the derivative zero, we get

$$Y^* = \lambda(I - (1-\lambda)V)^{-1}X, \tag{5.15}$$

where $V = D^{-1/2}WD^{-1/2}$.

5.2.4 Protein inference model

Based on the LR probability of peptides, the protein score is computed as,

$$s_k = \frac{1}{N_k}\sum_{i=1}^{n_k} q_i \tag{5.16}$$

where s_k is the score of protein Q_k and q_i is the LR probability of peptide P_i. n_k and N_k are the number of experimental and theoretical peptides for protein Q_k, respectively. The number of theoretical peptides N_k is included to factor the length of a protein in the model. It is computed based on these criteria: (1) trypsin-cutting; (2) two missed cleavages are allowed; and (3) peptides with masses falling in $[M_{min}, M_{max}]$. The minimum (M_{min}) and maximum (M_{max}) peptide masses are determined from the peptide identification reports. An alternative way is to only count peptides with a certain length [17].

5.2.5 Experimental Data

Two datasets are constructed and analyzed with the proposed method, and they were described in [26]. These datasets are generated on Thermo Electron (Waltham, MA) LTQ and ABI (Foster City, CA) API QSTAR Pulsar i, respectively. They are used to verify the robustness of our method to datasets from different experiments. To avoid high-dimension (over 10,000) matrix in the computation, we construct two

sub-datasets from the original datasets by keeping all the true peptides and randomly selecting similar number of false peptides. The summary of the two sub-datasets is given in Table 5.1.

Table 5.1: Statistics of the two sub-datasets. The number of true proteins including standard proteins and contaminant ones is given in the table. Besides, the number of true and false peptides in the constructed datasets and those which are also output from PeptideProphet with probability>0.05 (in brackets) are summarized as well.

	Standard Proteins	Contaminants	True peptides	False peptides	Peptides
Mix1	18	13	4318 [1218]	4610 [998]	8928 [2216]
Mix2	18	15	1689 [792]	3605 [408]	5294 [1200]

5.3 Results

The proposed method is compared with PeptideProphet [19] and ProteinProphet [1] for the peptide identification and protein inference, respectively. Specifically, receiver operating characteristic (ROC) curves are used to measure the discrimination power of the LR probability and PeptideProphet probability for peptide identification, and the coverage of identified proteins is compared for protein inference.

5.3.1 Parameters setting

The parameters in logistic regression are computed with the active learning method, the details of which can be referred to [22]. Here, $\lambda_1 = [5, 50, 50]^T$ and $\mu_0 = 0.0001[1, 1, 1]^T$. Since the variance of $\mathbf{w}_i$ is $\frac{1}{\lambda_{1i}}$, we allow $\mathbf{w}_0$ has the largest variance of $1/5$, while the other two parameters have the same variance of $1/50$.

5.3.2 Peptide identification results

We compute the LR probability with the original and regularized Sequest scores. The results are given in Figure 5.2. The results of both Mix1 and Mix2 show that the discrimination power of LR probability based on the original Sequest scores is much lower than LR probability based on regularized scores. Moreover, the best results are given by the scores regularized with the adjacency matrix (W_2) constructed from the selected high-confidence proteins. This indicates that the adjacency matrix updated with the selected high-confidence proteins can increase the confidence of peptides from high-confidence proteins while reduce the confidence of peptides from low-confidence proteins.

5.3.3 Comparison with PeptideProphet

We first show the number of identified peptides from the proposed feedback method and PeptideProphet, which is given in Table 5.2. By applying the feedback method to the two datasets, we can identify 3572 true peptides for Mix1 and 1511 true peptides for Mix2 given the false positive rate (FPR) around 0.05. At the same FPR, PeptideProphet can only identify 929 and 649 true peptides for Mix1 and Mix2, respectively. Furthermore, among the identified peptides by the feedback method, the numbers of peptides that are also output by PeptideProphet are shown in brackets. It can be seen that the proposed feedback method can identify much more true positive peptides while outputting much fewer false positive peptides than PeptideProphet.

To better compare the performance of the feedback method and PeptideProphet, Figures 5.3 shows the ROC curves of the two methods applied on the peptides output by PeptideProphet. It is very obvious that the feedback method has much higher discrimination power than PeptideProphet on these two datasets. This implies that the feedback from protein inference, i.e., the updated adjacency matrix between peptides, can essentially improve peptide scores, and thus increase the number of identified peptides.

Table 5.2: The number of identified peptides. By applying the feedback method on the two datasets, we can identify 3572 true peptides for Mix1 and 1511 true peptides for Mix2 given the false positive rate (FPR) around 0.05. At the same FPR, PeptideProphet can only identify 929 and 649 true peptides for Mix1 and Mix2, respectively. Furthermore, among the identified peptides by the feedback method, the numbers of peptides which are also output by PeptideProphet are shown in brackets. It can be seen that the proposed feedback method can identify much more true positive while output much fewer false positive peptides than PeptideProphet.

	Mix1		Mix2	
	Feedback	PeptideProphet	Feedback	PeptideProphet
True Positive	3572 [1193]	929	1511 [782]	649
False Positive	235 [3]	50	182 [4]	24
True Negative	4375 [995]	948	3423 [404]	384
False Negative	746 [25]	289	178 [9]	143

5.3.4 Protein inference results

Given FPR as 0.05, an LR probability threshold is determined and is used to filter peptides. Then, protein scores are computed as the sum of LR probability of the filtered constituent peptides. In the feedback workflow, high-confidence proteins are selected by setting an FPR of 5% as the threshold for protein inference. The coverage of these high-confidence proteins is then computed, and the final list of identified proteins is determined according to the protein coverage. First, we show the ROC curves of protein inference for Mix1 and Mix2 in Figure 5.4. It can be seen that the discrimination power of the feedback method is much higher than that of ProteinProphet. This is also illustrated in Table 5.3, which gives the number of inferred proteins of Mix1 and Mix2 at an FPR of 5%. It shows that not only can the feedback method infer more true positive proteins than ProteinProphet, but also output less false positive proteins than ProteinProphet. In addition, the coverage of the 33 true proteins is given in Table 5.6.

Table 5.3: The number of inferred proteins. The number of inferred proteins at FPR of 5% is shown. The feedback method not only can infer more true positive proteins than ProteinProphet, but also output less false positive proteins than ProteinProphet.

	Mix1		Mix2	
	Feedback	ProteinProphet	Feedback	ProteinProphet
True Positive	24	16	26	13
False Positive	4	38	4	19
True Negative	71	644	66	321
False Negative	9	15	7	17

5.4 Conclusion

We have demonstrated a new method to process peptide identification reports from database search engines. Protein inference and peptide identification are unified in this new method by adding a feedback from protein inference to peptide identification. The results have shown that the logistic regression based on scores that are regularized with the adjacency matrix has a much higher discrimination power than PeptideProphet. At the same FPR, our method can infer much more true positive proteins and less false positive proteins than ProteinProphet. In addition, the coverage of proteins inferred from the proposed method is much higher than the coverage computed from ProteinProphet. All these results indicate that the adjacency matrix between peptides which is constructed from the feedback of inferred proteins has an essential impact on the improvement of peptide scores.

Table 5.6: The coverage of true proteins. This table shows the coverage of 33 true proteins in the sample. For most true proteins, the proposed feedback method can significantly increase their coverage. The reason that some proteins have a coverage of 0 is because the peptides corresponding to these proteins have LR probability lower than the filter threshold. Similarly, the reason that the coverage is not available for some proteins from ProteinProphet is that the peptides input to ProteinProphet are filtered by PeptideProphet (probability> 0.05). It can be seen that the coverage of standard proteins in the sample is very high from the feedback method, and both methods can always identify peptides for these proteins, except ProteinProphet for protein *P02602*.

	Mix1		Mix2	
	Feedback	ProteinProphet	Feedback	ProteinProphet
>sp|P02188|MYG_HORSE	1	0.728	0.843	0.700
>sp|P02754|LACB_BOVIN	0.994	0.746	0.809	0.479
>sp|P46406|G3P_RABIT	0.988	0.804	0.849	0.645
>sp|Q29443|TRFE_BOVIN	0.987	0.684	0.790	0.630
>sp|P00722|BGAL_ECOLI	0.982	0.895	0.829	0.681
>sp|P00489|PHS2_RABIT	0.976	0.751	0.837	0.595
>sp|P00432|CATA_BOVIN	0.972	0.034	0.767	0.0884
>[Contaminant]sp|P02608|MLRS_RABIT	0.970	0.175	0.893	0.116
>sp|P00634|PPB_ECOLI	0.968	0.504	0.868	0.199
>sp|P06278|AMY_BACLI	0.965	0.741	0.650	0.637
>[Contaminant]sp|P01948|HBA_RABIT	0.964	0.477	0.354	0.088
>sp|P02602|MLE1_RABIT	0.963	0.033	0.874	0

continued on next page

continued from previous page

	Mix1		Mix2	
	Feedback	ProteinProphet	Feedback	ProteinProphet
>[Contaminant]sp|P02643|TNNI2_RABIT	0.961	0.0799	0.425	*N/A*
>sp|P02769|ALBU_BOVIN	0.960	0.758	0.837	0.618
>sp|P02666|CASB_BOVIN	0.960	0.438	0.754	0.518
>sp|P00921|CAH2_BOVIN	0.958	0.795	0.792	0.645
>sp|P00946|MANA_ECOLI	0.957	0.836	0.831	0.581
>sp|P01012|OVAL_CHICK	0.953	0.788	0.826	0.604
>[Contaminant]sp|P0AF93|YJGF_ECOLI	0.937	N/A	0.654	*N/A*
>sp|P00711|LCA_BOVIN	0.937	0.745	0.725	0.542
>[Contaminant]sp|P0A6F3|GLPK_ECOLI	0.934	N/A	0.110	*N/A*
>[Contaminant]sp|P02057|HBB_RABIT	0.932	0.654	0.815	0.173
>[Contaminant]sp|P02586|TNNC2_RABIT	0.931	0.109	0.553	0.035
>[Contaminant]sp|P58772|TPM1_RABIT	0.915	N/A	0.592	*N/A*
>sp|P62894|CYC_BOVIN	0.895	0.775	0.695	0.562
>[Contaminant]sp|P62975|UBIQ_RABIT	0.895	N/A	0	*N/A*
>sp|P62739|ACTA_BOVIN	0.641	0.860	0.589	0.739
>[Contaminant]sp|O46375|TTHY_BOVIN	0	0.610	0.517	0.466
>[Contaminant]sp|P81178|ALDH2_MESAU	0	0.491	0	0.170
>[Contaminant]sp|P00883|ALDOA_RABIT	0	0.893	0	0.503
>[Contaminant]sp|P01088|ITRF_MAIZE	0	0.392	0.574	0.116
>[Contaminant]sp|P69327|AMYG_ASPAW	0	N/A	0.648	*N/A*
>[Contaminant]sp|Q08043|ACTN3_HUMAN	0	N/A	0	*N/A*

Acknowledgment

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Conflict of interests

The authors declare that they have no competing financial/commercial interests.

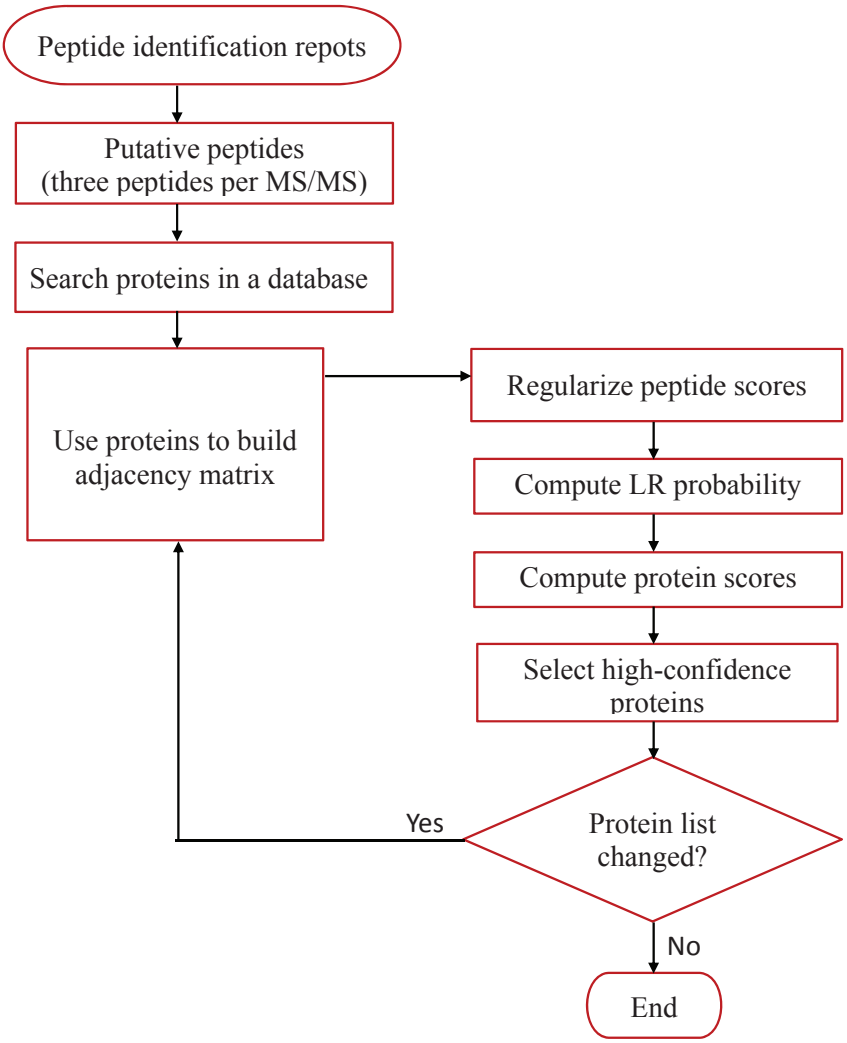


Figure 5.1: Feedback workflow for peptide identification and protein inference. The starting point are the peptide identification reports from database search engines. First, multiple peptides are selected for each MS/MS. Second, putative peptides are used to search proteins in the database. Third, an adjacency matrix which shows whether two peptides are siblings or not is built according to the list of proteins. Then, peptide scores are regularized with the application of two consistency assumptions, and the regularized scores are used as features in logistic regression (LR) to compute peptide identification probability. Based on the LR probability, protein scores are computed. Next, high-confidence proteins are selected to compose the new list of proteins, which is used to update the adjacency matrix between peptides. The experiments have shown that the loop will stop in two to four iterations for the used datasets.

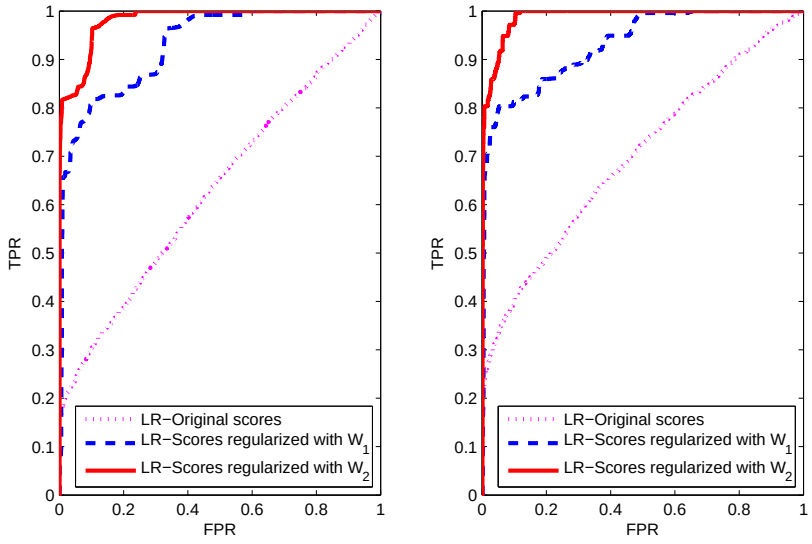


Figure 5.2: The results of Mix1 and Mix2 show that the discrimination power of LR probability based on the original Sequest scores is much lower than LR probability based on regularized scores. Moreover, the best results are given by the scores regularized with the adjacency matrix (W_2) constructed from the selected high-confidence proteins. This indicates that the adjacency matrix updated with the selected high-confidence proteins can increase the confidence of peptides from high-confidence proteins while reduce the confidence of peptides from low-confidence proteins.

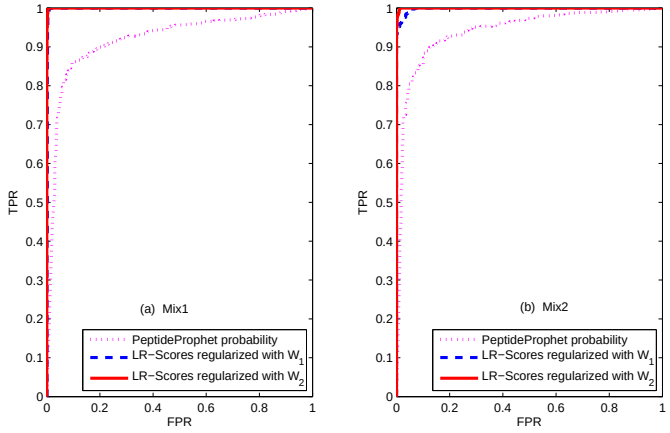


Figure 5.3: ROC curves show that the feedback method has much higher discrimination power than PeptideProphet on both datasets. This implies that the feedback from protein inference, i.e., the updated adjacency matrix between peptides, can essentially improve peptide scores, and thus increase the number of identified peptides.

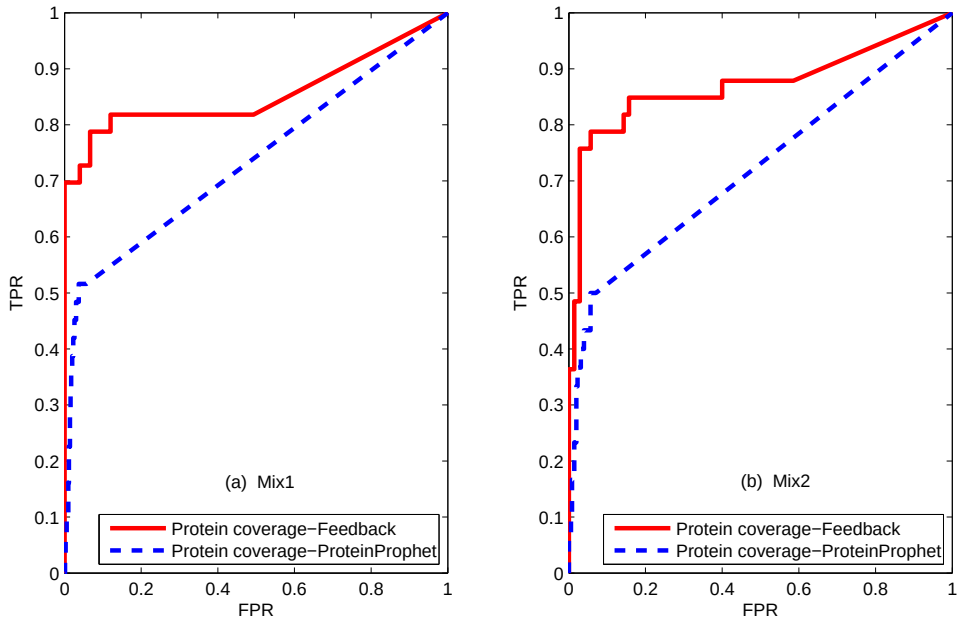


Figure 5.4: It can be seen that the discrimination power of the feedback method is much higher than that of ProteinProphet for both datasets.

References

[1] Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., "A statistical model for identifying proteins by tandem mass spectrometry," *Anal. Chem.*, 75: 4646-4658, 2003.

[2] Pieroni, E., Bentem, S., Mancosu, G., Capobianco, E., Hirt, H., and Fuente. A., "Protein networking: insights into global functional organization of proteomes", *Proteomics*, 8: 799-816, 2008.

[3] Hyungl, S. J., Ruotolo, B. T., "Integrating mass spectrometry of intact protein complexes into structural proteomics", *Proteomics*, 12: 1547-1564, 2012.

[4] Nesvizhskii, A. I. and Aebersold, R., "Interpretation of shotgun proteomic data: the protein inference problem," *Mol. Cell. Proteomics*, 4(10): 1419-1440, 2005.

[5] Shi, J., Wu, F.-X., "Protein inference by assembling peptides identified from tandem mass spectra," *Curr. Bioinform.*, 4: 226-233, 2009.

[6] T. Huang, J. Wang, W. Yu, and Z. He, "Protein inference: a review," *Brief. Bioinform.*, 13: 586-614, 2012.

[7] Price, T. S., Lucitt, M. B., Wu, W., Austin, D. J., Pizarro, A., Yocum, A. K., Blair, I. A., FitzGerald, G. A. and Grosser, T., "EBP: a program for protein identification using multiple tandem mass spectrometry datasets," *Mol. Cell. Proteomics*, 6: 527-536, 2007.

[8] Alves, P., Arnold, R. J., Novotny, M. V., Radivojac, P., Reilly, J. P. and Tang, H., "Advancement in protein inference from shotgun proteomics using peptide detectability," *Pac. Symp. Biocomput.*, 12: 409-420, 2007.

[9] Y. F. Li, R. J. Arnold, Y. Li, P. Radivojac, Q. Sheng and H. Tang, "A Bayesian approach to protein inference problem in shotgun proteomics," *J. Comput. Biol.*, 16:1183-1193, 2009.

[10] J. A. Taylor and R. S. Johnson, "Sequence database searches via de novo peptide sequencing by tandem mass spectrometry," *Rapid Commun. Mass Spectrom.*, 11: 1067-1075, 1997.

[11] Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A. and Lajoie, G., "PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry," *Rapid Commun. Mass Spectrom.*, 17: 2337-2342, 2003.

[12] Mo, L., Dutta, D., Wan, Y. and Chen, T., "MSNovo: A dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry," *Anal. Chem.*, 79: 4870-4878, 2007.

[13] Perkins, D. N., Pappin, D. J. C., Creasy, D. M., and Cottrell, J. S., "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, 20: 3551-3567 1999.

[14] Eng, J. K., McCormack, A. L., and Yates, J. R. III, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *J. Am. Soc. Mass Spectrom.*, 5: 976-989, 1994.

[15] Craig, R. and Beavis, R. C. "TANDEM: matching proteins with tandem mass spectra," *Bioinformatics*, 20: 1466-1467, 2004.

[16] Shen, C., Wang, Z., Shankar, G., Zhang, X. and Li, L., "A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry," *Bioinformatics*, 24: 202-208, 2007.

[17] Spivak, M., Tomazela, D., Weston, J., MacCoss, M. J., and Noble, W. S., "Direct maximization of protein identifications from tandem mass spectra," *Mol. Cell. Proteomics*, 2011.

[18] Li, Q., MacCoss, M. and Stephens, M., "A nested mixture model for protein identification using mass spectrometry," *Ann. Appl. Stat.*, 4(2): 962-987, 2010.

[19] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search," *Anal. Chem.*, 74: 5383-5392, 2002.

[20] Ding, Y., Choi, H., and Nesvizhskii, A. I., "Adaptive discriminant function analysis and reranking of MS/MS database search results for improved peptide identification in shotgun proteomics," *J. Proteome Res.*, 7: 4878-4889, 2008.

[21] Bishop, C. M., *Pattern Recognition and Machine Learning*, Springer, Singapore, 2006.

[22] Shi, J., Lin, W., Wu, F.-X., "Statistical analysis of Mascot peptide identification with active logistic regression," *iCBBE*, 2010.

[23] Boyd, S., Vandenberghe, L., Convex Optimization. Cambridge University Press, New York, 2004.

[24] He, Z., Zhao, H., Yu, W., "Score regularization for peptide identification," *BMC Bioinfomatics*, 12(Suppl):S2, 2011.

[25] D. Zhou, O. Bousquet, T. N. Lal, J. Weston and B. Schölkopf, "Learning with local and global consistency." *NIPS*, 2003.

[26] J. Klimek, J. S. Eddes, L. Hohmann, J. Jackson, A. Peterson, S. Letarte, P. R. Gafken, J. E. Katz, P. Mallick, H. Lee, A. Schmidt, R. Ossola, J. K. Eng, R. Aebersold and D. B. Martin, "The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools," *J. Proteome Res.*, 7: 96-103, 2008.

Chapter 6

Conclusions, contributions and recommendations

6.1 General discussion

Protein inference based on peptides identified from tandem mass spectra is an important computational step in the study of proteomics. The MS-based protein inference problem can be divided into three computational phases: (1) process MS/MS to improve the quality of the data and facilitate subsequent peptide identification; (2) postprocess peptide identification results from existing algorithms that match MS/MS to peptides; and (3) infer proteins by assembling identified peptides. The addressing of these computational problems has consisted of the main content of this thesis. In addition, the basic concepts and principles of mass spectrometry in proteomics were introduced, and the major strategies for peptide identification and protein inference were reviewed. In the following, a general discussion is given to summarize the relationship of each manuscript to the thesis and how they make the thesis as a whole.

The manuscript included in Chapter 3 studies the determination of low-resolution CID tandem mass spectra with an unsupervised machine learning method GMM. The determination of charge states of tandem mass spectra is an important aspect in the processing of MS/MS data before peptide identification, which belongs to the first computational phase of protein inference. It can be included as a preprocessing step in designing programs for peptide identification. In Chapter 4, the manuscript included proposes a new method to estimate the accuracy of peptide identification with logistic regression (LR) based on Sequest scores. This

is the necessary step to postprocess peptide identification results from search engines for several reasons. First, there is usually a high false positive rate in the results, which can bring many false identifications to protein inference. Secondly, the postprocessing makes it possible to compare and combine the results from different search engines, and also facilitates subsequent protein inference. As one step in the second computational phase, this manuscript is an important component part of the thesis. Finally, the manuscript included in Chapter 5 proposes an iterative method based on a unified framework to infer proteins with peptides identified from Sequest. The statistical analysis of peptide identification is performed with the LR introduced in Chapter 4. Protein inference and peptide identification are iterated in one framework by adding a feedback from protein inference to peptide identification. This is the last computational step in MS-based protein inference. To summarize, the three manuscripts included in this thesis are all closely related to the thesis topic and make the thesis as a whole.

6.2 Summary of conclusions, contributions and recommendations

Based on this research, the following conclusions are drawn as such:

- Gaussian mixture model with novel and discriminant features can accurately determine the charge states of low-resolution CID peptide tandem mass spectra. Especially, the newly proposed feature $\delta_{\mathrm{R_{cp}}}$, which measures the difference between the ratio of +1 peak intensity over their complementary + 1 peak intensity and the ratio between +2 peak intensity over their complementary +1 peak intensity, is the most discriminant feature in determining the charge states.

- Logistic regression is an easy and effective model to compute the peptide identification probability based on regularized search engine scores.

- The adjacency matrix between peptides is a significant factor in improving peptide identification accuracy, because it captures extra protein information into the peptide identification step.

- It is advantageous to perform protein inference and peptide identification under the proposed unified framework, which not only can improve the accuracy of peptide identification, but also can increase the accuracy and coverage of protein inference.

The major contributions of the research can be summarized as follows:

- The review of the mainstream methods of addressing the protein inference problem and the major challenges arisen from the problem is provided.

- An easy machine learning method (GMM) is proposed to determine the charge states of low-resolution CID peptide tandem mass spectra with four novel and highly discriminant features to represent each spectrum. This unsupervised method is especially useful when the training data is expensive to collect or not available.

- The statistical analysis of peptide identification results from search engines such as Sequest has been conducted. This work is necessary due to the large scale of MS data which makes it impractical to manually verify the identification results. In addition, the statistical analysis can unify identification results from different search engines into the same scale, which can be used to compare and, more importantly, to combine these results for the subsequent protein inference. Statistical analysis results can be easier to use directly in protein inference models. The generated probability is a comprehensive reflection of all main factors considered in peptide identification.

- A unified framework and an iterative method are developed to infer proteins and identify peptides simultaneously based on the peptide identification reports from search engines. The key point in this framework is to update the adjacency matrix between peptides by use of the feedback of a list of high-confidence proteins from protein inference to peptide identification. The adjacency matrix is used to regularize peptide scores, the results of which are used in a logistic regression model to compute probabilities of peptide identifications. This can greatly improve the accuracy of peptide identification,

because this adjacency matrix captures extra protein information into peptide identification. Besides, multiple peptides are selected for each tandem mass spectrum, and these peptides are given a second chance to be reevaluated. With the enhancement obtained from the regularization of scores, many more peptides are identified, and they further contribute to the inference of proteins in terms of increasing the accuracy and coverage of inferred proteins.

Some future work is recommended:

- The assignment of degenerate peptides to truly present parent proteins is still a challenge in protein inference. The proposed MS/MS intensity-based strategy of assigning degenerate peptides is tentative. Improvements can be made, such as, in the computation of peptide intensity.

- The proposed methods are verified on relatively simple datasets, which are collected for the verification of proteomics algorithms. There shouldn't be any scalability problem of the proposed unified framework of protein inference, as long as the operations (multiplication and inverse) of large matrices is not a problem. So it can be tested on more complex datasets to be further verified.

- It is still a challenge to validate the protein inference results in proteomics. It could be an independent research topic which deserves more efforts from researchers.

- As more and more supplementary information becomes available, protein inference can be performed by combing this information with traditional MS data. The supplementary information such as, raw MS/MS spectra, single-stage MS data, peptide expression profiles, mRNA expression data, PPI networks or gene models, can be used to address the ambiguity problem in protein inference brought by the degenerate peptides and 'one-hit wonders'.

Appendix A

Publications

Journal

[1] **J. Shi**, B. Chen and F. X. Wu, "Unifying protein inference and peptide identification with feedback to update consistency between peptides," *Proteomics*, 2: 1-9, 2012.

[2] **J. Shi**, and F. X. Wu, "A feedback framework for protein inference with peptides identified from tandem mass spectra," *Proteome Science*, 2012, accepted.

[3] Z. Yuan, **J. Shi**, W. Lin, B. Chen and F. X. Wu, "Features-based deisotoping method for tandem mass spectra," *Advances in Bioinformatics*, Volume 2011, 12 pages.

[4] **J. Shi**, and F. X. Wu, "Peptide charge state determination of tandem mass spectra from low-resolution collision induced dissociation," *Proteome Science*, 9(Suppl1):S3, 2011.

[5] W. Lin, F. X. Wu, **J. Shi**, J. Ding, and W. Zhang, "An adaptive approach to denoising tandem mass spectra," *Proteomics*, 11: 3773-3778, 2011.

[6] J. Ding, **J. Shi**, and F. X. Wu, "SVM-RFE based feature selection for tandem mass spectrum quality assessment," *International Journal of Data Mining and Bioinformatics*, 5(1): 73-88, 2011.

[7] A.M. Zou, **J. Shi**, J. Ding and F. X. Wu, "Charge state determination of peptide tandem mass spectra using support vector machine (SVM)," *IEEE Transaction on Information Technology in Biomedicine*, 14(3): 552-558, 2010.

[8] **J. Shi** and F. X. Wu, "Protein inference by assembling peptides identified from tandem mass spectra," *Current Bioinformatics*, 4(3):226-233, 2009.

[9] J. Ding, **J. Shi**, and F. X. Wu, "Quality assessment of tandem mass spectra by using a weighted k-means," *Clinical Proteomics*, 5(1): 15-22, 2009.

[10] J. Ding, **J. Shi**, G. G. Poirier, and F. X. Wu, "A novel approach to denoising ion trap tandem mass spectra," *BMC Proteome Science*, 7:9, 2009.

Conference

[1] **J. Shi**, B. Chen and F. X. Wu, "Improve accuracy of peptide identification with consistency between peptides," *IEEE BIBM'2011*, 191-196.

[2] B. Chen, **J. Shi**, Y. Yan, S. Zhang and F. X. Wu, "An improved graph entropy-based method for identifying protein complexes," *IEEE BIBM'2011*, 123-126.

[5] **J. Shi**, W. Lin and F. X. Wu, "Statistical analysis of Mascot peptide identification with active logistic regression," *iCBBE2010*, 1-4.

[6] J. Ding, **J. Shi**, and F. X. Wu, "Model based clustering for tandem mass spectrum quality assessment," *IEEE EMBC'2009*, 6747-6750.

[7] J. Ding, **J. Shi**, A.M. Zou, and F. X. Wu, "Feature selection for tandem mass spectrum quality assessment," *IEEE BIBM'2008*, 310-313.

Book chapter

[1] **J. Shi**, and F. X. Wu, "Assigning probabilities to Mascot peptide identification using logistic regression," Advances in Experimental Medicine and Biology, 1, Volume 680, Advances in Computational Biology, Part 3: 229-236.

Appendix B

Copyright permissions

The copyright of the following papers:

Jinhong Shi, and Fang-Xiang Wu, "Protein inference by assembling peptides identified from tandem mass spectra," *Current Bioinformatics*, 4(3):226-233, 2009.

Jinhong Shi, and Fang-Xiang Wu, "A feedback framework for protein inference with peptides identified from tandem mass spectra," *Proteome Science*, 2012, accepted.

Jinhong Shi, and Fang-Xiang Wu, Peptide charge state determination of tandem mass spectra from low-resolution collision induced dissociation, *Proteome Science*, vol.9(Suppl 1):S3, 2011.

Jinhong Shi, Bolin Chen and Fang-Xiang Wu, "Improve accuracy of peptide identification with consistency between peptides," *IEEE BIBM'2011*, Atlanta, America, 12-15 November 2011.

Jinhong Shi, Bolin Chen and Fang-Xiang Wu, "Unifying protein inference and peptide identification with feedback to update consistency between peptides," *Proteomics*, 2: 1-9, 2012.

are included in the following pages.

Grant of Permission

AMBREEN IRSHAD - BSP [ambreenirshad@benthamscience.org]
Sent: Tuesday, November 06, 2012 10:03 PM
To: Shi, Jinhong
Cc: m.ahmed@benthamscience.org

Grant of Permission

Dear Dr. Jinhong:

Thank you for your interest in our copyrighted material, and for requesting permission for its use.

Permission is granted for the following subject to the conditions outlined below:

Jinhong Shi, and Fang-Xiang Wu, Protein Inference by Assembling Peptides Identified from Tandem Mass Spectra," Current Bioinformatics, 4(3):226-233, 2009

To be used in the following manner:

1. Bentham Science Publishers grants you the right to reproduce the material indicated above on a one-time, non-exclusive basis, solely for the purpose described. Permission must be requested separately for any future or additional use.

2. For an article, the copyright notice must be printed on the first page of article or book chapter. For figures, photographs, covers, or tables, the notice may appear with the material, in a footnote, or in the reference list.

Thank you for your patience while your request was being processed. If you wish to contact us further, please use the address below.

Sincerely,

AMBREEN IRSHAD
Permissions & Rights Manager
Bentham Science Publishers
Email: permission@benthamscience.org
URL: www.benthamscience.com

From: Jinhong Shi [mailto:jis958@mail.usask.ca]
Sent: Wednesday, November 07, 2012 3:16 AM
To: permission@benthamscience.org
Subject: request permission for reuse of paper in doctoral dissertation

Dear Sir/Madam:

I am completing a manuscript-based doctoral dissertation at University of Saskatchewan. I would like your permission to reuse in my dissertation the following paper manuscript:

Jinhong Shi, and Fang-Xiang Wu, Protein Inference by Assembling Peptides Identified from Tandem Mass Spectra,'' Current Bioinformatics, 4(3):226-233, 2009.

The contents to be reused are the whole paper manuscript. Would you please give me a reply by Nov. 14, 2012? Thank you very much.

Jinhong

Copyright

Research articles

Copyright on any research article in a journal published by BioMed Central is retained by the author(s).

Authors grant BioMed Central a license to publish the article and identify itself as the original publisher.

Authors also grant any third party the right to use the article freely as long as its integrity is maintained and its original authors, citation details and publisher are identified.

The BioMed Central Copyright and License Agreement (identical to the Creative Commons Attribution License) formalizes these and other terms and conditions of publishing research articles.

Other articles

In general, authors retain copyright for any article that is published with open access but are asked to assign copyright to the publisher for articles that are not.

Authors' certification

In submitting a research article ('article') to any of the journals published by BioMed Central Ltd ('BioMed Central') authors are requested to certify that:

They are authorized by their co-authors to enter into these arrangements.

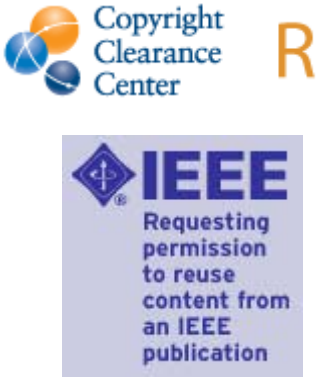
They warrant, on behalf of themselves and their co-authors, that:

the article is original, has not been formally published in any other peer-reviewed journal, is not under consideration by any other journal and does not infringe any existing copyright or any other third party rights;

they are the sole author(s) of the article and have full authority to enter into this agreement and in granting rights to BioMed Central are not in breach of any other obligation. If the law requires that the article be published in the public domain, they will notify BioMed Central at the time of submission;

the article contains nothing that is unlawful, libellous, or which would, if published, constitute a breach of contract or of confidence or of commitment given to secrecy;

they have taken due care to ensure the integrity of the article. To their - and currently accepted scientific - knowledge all statements contained in it purporting to be facts are true and any formula or instruction contained in the article will not, if followed accurately, cause any injury, illness or damage to the user.



Title:	Improve Accuracy of Peptide Identification with Consistency between Peptides
Conference Proceedings:	Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on
Author:	Jinhong Shi; Bolin Chen; Fang-Xiang Wu
Publisher:	IEEE
Date:	12-15 Nov. 2011

Logged in as:
Jinhong Shi

LOGOUT

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Nov 06, 2012

This is a License Agreement between Jinhong Shi ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3023241098222
License date	Nov 06, 2012
Licensed content publisher	John Wiley and Sons
Licensed content publication	Proteomics
Book title	
Licensed content author	Jinhong Shi,Bolin Chen,Fang-Xiang Wu
Licensed content date	Oct 30, 2012
Start page	n/a
End page	n/a
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Order reference number	
Total	0.00 USD
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a "Wiley Company") or a society for whom a Wiley Company has exclusive publishing rights in relation to a particular journal (collectively WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your Rightslink account (these are available at any time at http://myaccount.copyright.com)

Terms and Conditions

1. The materials you have requested permission to reproduce (the "Materials") are protected by copyright.

2. You are hereby granted a personal, non-exclusive, non-sublicensable, non-transferable, worldwide, limited license to reproduce the Materials for the purpose specified in the licensing process. This license is for a one-time use only with a maximum distribution equal to the number that you identified in the licensing process. Any form of republication granted by this licence must

be completed within two years of the date of the grant of this licence (although copies prepared before may be distributed thereafter). The Materials shall not be used in any other manner or for any other purpose. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Material. Any third party material is expressly excluded from this permission.

3. With respect to the Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Materials without the prior permission of the respective copyright owner. You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Materials, or any of the rights granted to you hereunder to any other person.

4. The Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc or one of its related companies (WILEY) or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto.

5. NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

6. WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

7. You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

8. IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

9. Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

10. The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not

operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

11. This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

12. Any fee required for this permission shall be non-refundable after thirty (30) days from receipt.

13. These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

14. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

15. WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

16. This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

17. This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

Wiley Open Access Terms and Conditions

All research articles published in Wiley Open Access journals are fully open access: immediately freely available to read, download and share. Articles are published under the terms of the Creative Commons Attribution Non Commercial License. which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes. The license is subject to the Wiley Open Access terms and conditions: Wiley Open Access articles are protected by copyright and are posted to repositories and websites in accordance with the terms of the Creative Commons Attribution Non Commercial License. At the time of deposit, Wiley Open Access articles include all changes made during peer review, copyediting, and publishing. Repositories and websites that host the article are responsible for incorporating any publisher-supplied amendments or retractions issued subsequently.
Wiley Open Access articles are also available without charge on Wiley's publishing platform, **Wiley Online Library** or any successor sites.

Use by non-commercial users

For non-commercial and non-promotional purposes individual users may access, download, copy, display and redistribute to colleagues Wiley Open Access articles, as well as adapt, translate, text- and data-mine the content subject to the following conditions:

- The authors' moral rights are not compromised. These rights include the right of "paternity" (also known as "attribution" - the right for the author to be identified as such) and "integrity" (the right for the author not to have the work altered in such a way that the author's reputation or integrity may be impugned).
- Where content in the article is identified as belonging to a third party, it is the obligation of the user to ensure that any reuse complies with the copyright policies of the owner of that

content.

- If article content is copied, downloaded or otherwise reused for non-commercial research and education purposes, a link to the appropriate bibliographic citation (authors, journal, article title, volume, issue, page numbers, DOI and the link to the definitive published version on Wiley Online Library) should be maintained. Copyright notices and disclaimers must not be deleted.
- Any translations, for which a prior translation agreement with Wiley has not been agreed, must prominently display the statement: "This is an unofficial translation of an article that appeared in a Wiley publication. The publisher has not endorsed this translation."

Use by commercial "for-profit" organisations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee. Commercial purposes include:

- Copying or downloading of articles, or linking to such articles for further redistribution, sale or licensing;
- Copying, downloading or posting by a site or service that incorporates advertising with such content;
- The inclusion or incorporation of article content in other works or services (other than normal quotations with an appropriate citation) that is then available for sale or licensing, for a fee (for example, a compilation produced for marketing purposes, inclusion in a sales pack)
- Use of article content (other than normal quotations with appropriate citation) by for-profit organisations for promotional purposes
- Linking to article content in e-mails redistributed for promotional, marketing or educational purposes;
- Use for the purposes of monetary reward by means of sale, resale, licence, loan, transfer or other form of commercial exploitation such as marketing products
- Print reprints of Wiley Open Access articles can be purchased from: corporatesales@wiley.com

Other Terms and Conditions:

BY CLICKING ON THE "I AGREE..." BOX, YOU ACKNOWLEDGE THAT YOU HAVE READ AND FULLY UNDERSTAND EACH OF THE SECTIONS OF AND PROVISIONS SET FORTH IN THIS AGREEMENT AND THAT YOU ARE IN AGREEMENT WITH AND ARE WILLING TO ACCEPT ALL OF YOUR OBLIGATIONS AS SET FORTH IN THIS AGREEMENT.

v1.7

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK500892079.
Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center

Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.